VOL FOUR / ISSUE FOUR

NewScientist THE COLLECTION

INFINITY AND BEYOND

YOUR ULTIMATE GUIDE TO MATHEMATICS

ZERO, INFINITY, WONDER NUMBERS, PROBABILITY AND STATISTICS, SYMMETRY AND REALITY, COMPUTATIONAL COMPLEXITY, MATHS AND YOUR BRAIN

£9.99

9 772054 638003 04

invitrogen

Streamline analysis, expand possibilities

E-Gel Power Snap Electrophoresis System

DNA separation

Simplify DNA electrophoresis with the only integrated gel running and imaging platform

The new Invitrogen™ E-Gel™ Power Snap Electrophoresis System combines the convenience of rapid, real-time nucleic acid analysis with high-resolution image capture. The integrated design helps reduce workflow time and accelerate discovery.

Find out more at **thermofisher.com/powersnap**

ThermoFisher SCIENTIFIC

NewScientist
THE COLLECTION

VOL FOUR / ISSUE FOUR
INFINITY AND BEYOND

NEW SCIENTIST
THE COLLECTION
25 Bedford Street,
London WC2E 9ES
+44 (0)20 7611 1202
enquiries@newscientist.com

Editor-in-chief Graham Lawton
Editor Richard Webb
Art editor Ryan Wills
Picture editor
New Scientist Picture Desk
Subeditor Hannah Joshua
Graphics Dave Johnston
Production editor Mick O'Hare
Project manager Henry Gomm
Publisher John MacFarlane

New Scientist The Collection is published four times per year by New Scientist Ltd
ISSN 2054-6386

Printed in England by William Gibbons (Wolverhampton) and distributed by Marketforce UK Ltd
+44(0)20 3148 3333

Display advertising
+44(0)20 7611 1291
displayads@newscientist.com

Cover image
Login/Shutterstock

Learn the language of reality

LOVE it or loathe it, there is no escape: mathematics is out there. We live in a world that is in some sense mathematical – although in precisely what sense is still hotly debated by mathematicians, physicists, philosophers and others. We are, as a result, innately mathematical beings. Crossing the road, catching a ball, stacking the dishwasher: in our everyday lives we are constantly, unconsciously, manipulating numbers, assessing shapes, calculating position, geometry and motion. We are doing maths.

But most of us see the conscious pursuit of the subject as something best left to the experts. The abstractions, conjectures and proofs of formal mathematics belong to a higher, rarefied plane few of us can access.

This latest issue of *New Scientist: The Collection* aims to bridge the divide between the sublime and the mundane, and explore both the mystery and fascination of figures. We start in Chapter 1 by asking what mathematics is and how it relates to us: why its practitioners are often bewitched by its beauty, while for others it simply does not add up.

In Chapter 2, it's time to pitch into the bread and butter of mathematics: numbers. Primes are the atoms of the number system, and attempts to understand how they work are essential to number theory and mathematics as a whole. But we look also at other figures of peculiar significance, from Euler's number, *e*, to the imaginary unit i – a number that shouldn't exist, but clearly does.

Similar conceptual difficulties surround the two bookends of the number system: zero and infinity. It took a long time for mathematicians to realise zero was a number at all – let alone that it has a claim to be the only number in existence. Its story is told in Chapter 3.

Infinity, meanwhile, is a monstrosity of a concept, too big for our brains, but it is the key that lets us coax mathematical logic into making any sense at all. That said, can it exist in the real world? Find out in Chapter 4.

Chapter 5 gets practical. The old chestnut "Lies, damn lies and statistics" is never truer than when those statistics apply to our health. We share some expert tips on how to avoid the wool being over your eyes – and on other mathematics essential for a smarter life.

Probability underlies statistics, and presents many a pitfall, even for experienced practitioners. Chapter 6 deals with the raging debates about what type of probability is best and how and where to apply it, as well as the difficulty of achieving true randomness.

In Chapter 7, it's computational complexity, the problem that would solve all other problems – if only we could solve it. That's a prelude to some more mundane maths in Chapter 8, explaining why voting systems are never fair, how best to slice a pizza and how to beat the bookies at their own game.

Chapter 9 changes gear once again, to address that most profound question of how mathematics truly relates to reality. We look at symmetry, the principle that makes and breaks the universe, and how the work of one largely forgotten woman a century ago brought it to the fore. Then there's the contention that mathematics doesn't just describe reality, it *is* reality – and the question of the degree to which randomness rules it.

Heady stuff, so to round off, it's a bit of miscellaneous fun: from efforts to boil down pasta shapes to a few bare formulae, to the true answer to life, the universe and everything, via the maths of *Alice in Wonderland* and the world's hardest logic puzzle.

So be challenged – and enjoy!

Richard Webb, Editor

CONTENTS

VOL 4 / ISSUE 4

INFINITY AND BEYOND

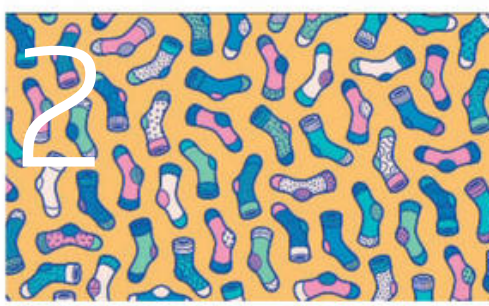
CONTRIBUTORS

Gilead Amit
is a feature editor at *New Scientist*
Anil Ananthaswamy
is a consultant for *New Scientist*
Jacob Aron
is analysis editor at *New Scientist*
Melanie Bayley
was a doctoral student at the University of Oxford and is a home tutor
Michael Brooks
is a consultant for *New Scientist*
Matthew Chalmers
is editor of *CERN Courier*
Stuart Clark
is a consultant for *New Scientist*
Daniel Cossins
is a feature editor at *New Scientist*
Richard Elwes
is a mathematician at the University of Leeds, UK
Marianne Freiberger
is co-editor of online maths magazine *Plus*
Amanda Gefter
is a writer based in Boston, Massachusetts
Dave Goldberg
is a physicist at Drexel University in Philadelphia
Christopher Kemp
is a writer based in Grand Rapids, Michigan
Stephen Ornes
is a writer based in Nashville, Tennessee
Timothy Revell
is a reporter at *New Scientist*
Angela Saini
is a writer based in London
Marcus du Sautoy
is a mathematician at the University of Oxford
Laura Spinney
is a writer based in Paris
Ian Stewart
is emeritus professor of mathematics at the University of Warwick, UK
Manya Raman Sundström
is a researcher at Umeå University, Sweden
Max Tegmark
is a physicist at the Massachusetts Institute of Technology in Boston
Rachel Thomas
is co-editor of online maths magazine *Plus*
Helen Thomson
is a consultant for *New Scientist*
Richard Webb
is chief features editor at *New Scientist*

The articles here were first published in *New Scientist* between September 2007 and October 2017. They have been updated and revised.

What is maths?

Wonder numbers



Zero

Infinity

Lies, damn lies and...

Probability

Computation

Everyday maths

Maths and reality

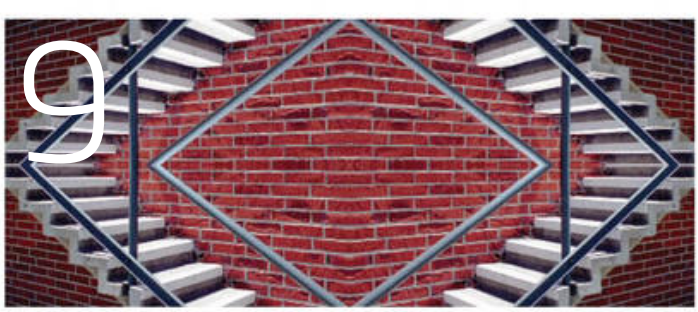
A mathematical miscellany

6

7

8

9



10

Commercialization | Technologies

IMS Health & Quintiles are now

IQVIA™

CHRISTIAN GRALINGEN

THE ORIGIN OF MATHEMATICS

It's our most effective tool for understanding the universe. But where it comes from and how it developed remain mysterious, finds Anil Ananthaswamy

TO THE Iranian mathematician Maryam Mirzakhani, the first woman to win the Fields medal, mathematics often felt like "being lost in a jungle and trying to use all the knowledge that you can gather to come up with some new tricks".

"With some luck," she added, "you might find a way out."

Mirzakhani, who died in July 2017 at the age of 40, ventured deeper into the mathematical jungle than most. Nonetheless, most of us have spent enough time on its periphery to have a sense of what the terrain looks like.

Increasingly, it seems as if humans are the only animals with the cognitive ability to hack their way through the undergrowth. But where does this ability come from? Why did we develop it? And what is it for? Answering these questions involves diving into one of the hottest debates in neuroscience, and reimagining what mathematics really is.

The natural world is a complex and unpredictable place. Habitats change, predators strike, food runs out. An organism's survival depends on its ability to make sense of its surroundings, whether by counting down to nightfall, figuring out the quickest way to escape danger or weighing up the spots most likely to have food. And that, says Karl Friston, a computational neuroscientist and physicist at University College London, means doing mathematics.

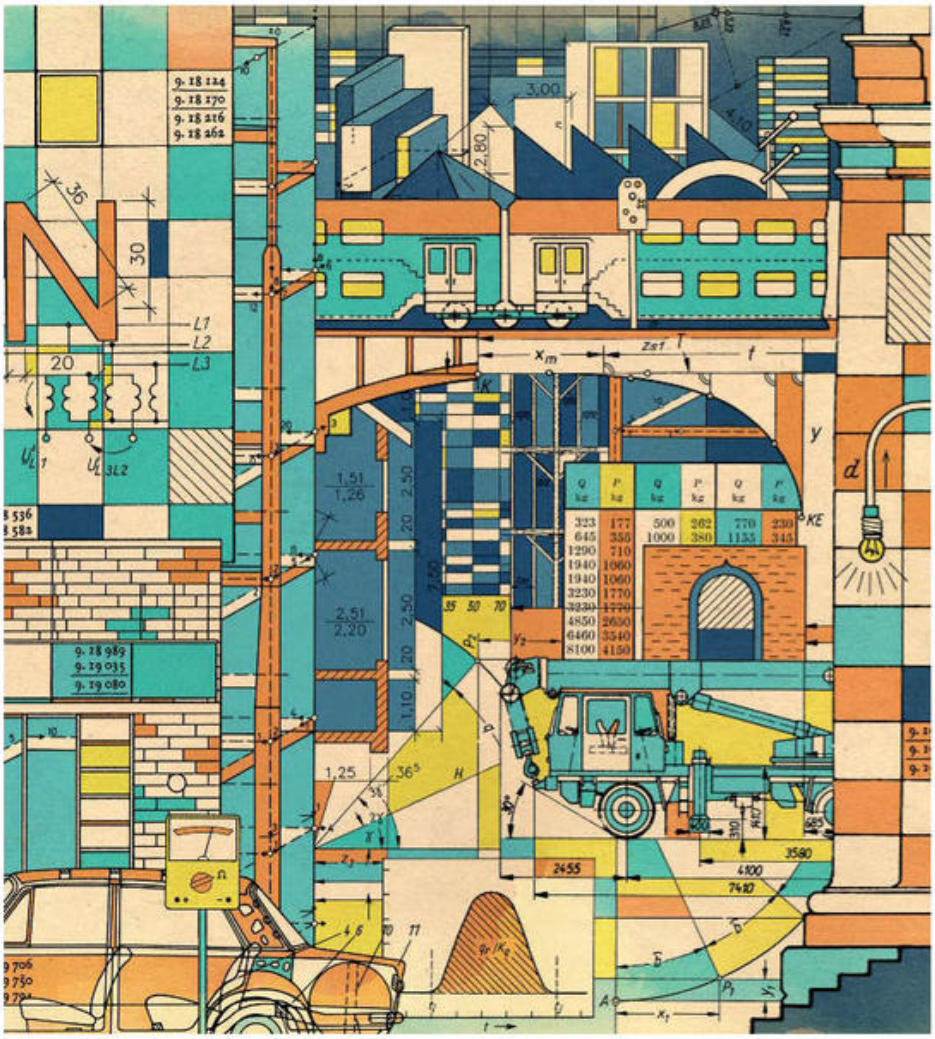
"There is a simplicity and parsimony and symmetry to mathematics," says Friston, "which, if you were treating it as a language, wins hands down over all other ways of describing the world." From dolphins to slime moulds, organisms throughout the evolutionary tree seem to make sense of the world mathematically, deciphering its patterns and regularities in order to survive.

Friston argues that any self-organising system – and so any form of life – that interacts with its environment needs an implicit model of that environment to function. The idea goes back to the 1970s and the "good regulator" theorem, co-developed by Ross Ashby, who pioneered the field of cybernetics. To provide effective control, the theorem says, a robot's brain must have an internal model of its mechanical body and its environment. "That insight is becoming increasingly formalised now in machine learning and artificial intelligence," says Friston. The corollary being that an animal's brain, too, must model its body and the world in which it moves.

No thought required

The remarkable thing is that none of these creature modellers are aware of what they're doing. Even we human beings, when we run to catch a ball or dart through heavy traffic, are unconsciously doing some pretty complex mathematics. Each of our brains is constantly using its models to predict what we will encounter, says the theory, and these models are kept updated by checking the predictions against actual sensations.

Those mathematical functions are undoubtedly being computed by particular bits of the brain, says Andy Clark, a cognitive philosopher at the University of Edinburgh, ›



UK. But this is not to say that there are specialised modules in the brain similar to buttons on a calculator that we can call up on demand: one to perform multiplication and another to work out cosines. "We don't have access to that," he says.

Although these models try to ensure our survival in a complex world that follows the laws of physics, their insistence on keeping us alive means they sometimes have to compromise on correctness. Take the gambler's fallacy: the mistaken belief that, if the roulette ball keeps landing on red, a bet on black is the best one to make. In reality, of course, both results are equally likely, but the models our brains have built of the world, perhaps to tell our ancestors when to move on from an unsuccessful foraging area, blind us to that simple statistical observation.

Or take the Weber-Fechner effect, which governs our response to external stimuli. Found to hold true across all our senses, it states that our ability to discriminate between sensations of a similar magnitude diminishes as their magnitudes increase together. So while a 1-kilogram weight can easily be distinguished from a 2-kilogram one, for example, weights of 21 kilograms and 22 kilograms are harder to tell apart. The same applies to the brightness of lights, the volume of sounds and even the number of objects you can see.

Although human brains share such aberrations with those of other animals, we have developed the ability to identify and overcome some of these flaws. Most obviously, we invented numbers: a system of notation that allows us to instantly deduce that 21 and 22 are as far apart as are 1 and 2. The creation of this complex, symbolic language for mathematics not only allows us to overcome certain such limitations of our subconscious mind, but also to explore abstract concepts in depth and communicate them to others. But how did we develop the tools to consciously understand what our bodies do instinctively?

> "We could have a sense of number as strong as our sense for colour"

One long-standing idea says we are born with a conscious sense of numbers in the same way we are conscious of colours. In his 1997 book *The Number Sense*, Stanislas Dehaene of the INSERM-CEA unit for cognitive neuroimaging in Gif-sur-Yvette, France, hypothesised that evolution endowed humans and other animals with numerosity, an ability to immediately perceive the number of objects in a pile. In other words, three red marbles would produce a sense of the number 3 just as they would produce a sense of the colour red. Dehaene proposed that this numerosity was exact for numbers below 4 and fuzzier thereafter, but nonetheless represented a hardwired ability. Armed with such an instinct, our paths through the mathematical jungle would quickly start to clear.

Innate numeracy

Evidence to support this "nativist" view soon started to accumulate. Elizabeth Spelke at the Massachusetts Institute of Technology and her colleagues showed that 6-month-old children could distinguish between an array of eight dots and one with 16 dots. Then Dehaene and his colleagues reported that the Munduruku Indians in the Brazilian Amazon, who don't have words for numbers larger than 5, could approximately discriminate between much larger quantities, suggesting that this ability was independent of culture.

Other studies indicated that humans instinctively represent numbers spatially on an imaginary "number line", their values increasing from left to right. There was even evidence of numerosity in animals (see "Animal instincts", page 10). This all pointed to an innate number sense that millennia of culture had helped expand.

But before long, some researchers grew uncomfortable with the conclusions of these studies. Might subjects, for example, be distinguishing two arrays of dots based not on the number of dots, but on other attributes such as their spatial distribution or area of coverage? "These are cues that are usually correlated with number, so it would be unwise not to use them," says Tali Leibovich at the University of Haifa in Israel. "If you are an animal in nature and you need to hunt something and need to do it very quickly, you want to use all available cues."

Indeed, on further examination, it seems that people also rely on these non-numerical

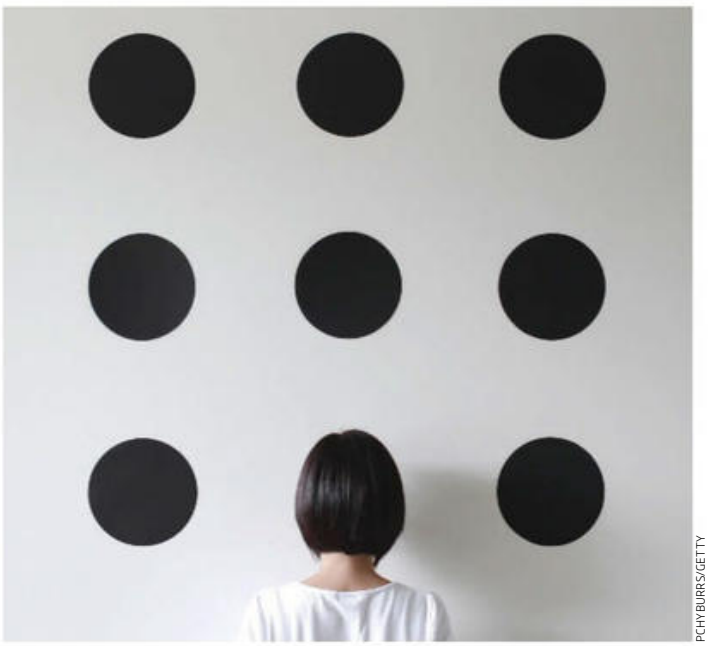
cues. Soon, a different hypothesis emerged. Perhaps, instead of having an innate sense of numbers, we are born with a sense for quantities – such as size and density – that correlate with the numbers of things. "It takes time and experience to develop and understand this correlation," says Leibovich.

More-refined cognitive tests in children tend to support this view. For example, children younger than about 4 years of age cannot understand that five oranges and five watermelons have something in common: the number 5. To them, a bunch of watermelons simply represents more "stuff" than the same number of oranges.

Even teaching young children to identify the order of numbers – going through the motions of counting – doesn't immediately impart their meaning, says developmental psychologist Daniel Ansari at the University of Western Ontario in Canada. That occurs informally through long-term exposure to parents and siblings. "This points to the strong influence of cultural practices on the learning of exact representations of number," he says.

Study of the cultural aspects of numerical cognition has suffered from bias, says Ansari, in that not enough attention has been paid to data collected from non-industrialised cultures. These findings, he believes, cast serious doubts on the nativist hypothesis.

Take the Yupno people of Papua New Guinea. Rafael Núñez at the University of California at San Diego has learned, for example, that they don't use the supposedly universal mental number line. Also, they have no comparatives in their language to say that one thing is bigger or smaller than another.



PCHYBURR/GETTY

Instinct or culture: How we grasp numbers is not all black and white

This is not to say that the Yupno language is primitive. Far from it. Take demonstratives. In English, there are only four: this, that, these and those, to specify the proximal or distal nature of things. The Yupno, on the other hand, have words to indicate whether something is higher or lower than them in elevation (in keeping with their mountainous terrain), and they have nuanced words to capture not only whether something is proximal or distal, but also by how much.

The Yupno are not alone in having a language that doesn't emphasise numbers. Núñez points to a study of 189 Aboriginal Australian languages, of which three-quarters were found to have no words for numbers above 3 or 4, while a further 21 went no higher than 5. To Núñez, this suggests that exact numerosity is a cultural trait that emerges when circumstances, such as agriculture and trading, demand it. "Hundreds of thousands of humans who have language, sometimes very complicated and sophisticated language, don't have exact quantification," he says.

Even languages that do, such as English or French, can only take you so far. In 2016, Dehaene and his student Marie Amalric reported the results of scanning the brains of 15 professional mathematicians and 15 non-mathematicians of the same academic standing. They found a network of brain regions involved in mathematical thought that was activated when mathematicians reflected on problems in algebra, geometry and topology, but not when they were thinking about non-mathsy things. No such distinction was visible in the other academics. Crucially, this "maths network" doesn't overlap with brain regions involved in language.

This suggests that once mathematicians have learned their symbolic language, they start thinking in ways that don't involve normal language. "It sounds strange, but it's almost like being able to download an intuition into another world, the world of mathematics, stand back, and let it talk back to you again," says Friston (see "Why do people hate maths?", page 11). ›

THE PILLARS OF MATHEMATICS

For most of us, maths means numbers, and that's not wrong. The ability to understand and manipulate numbers in the abstract (think addition, subtraction, multiplication and division) is the foundation on which a formidable edifice has been built (see main story). Broadly speaking, this edifice consists of three pillars: geometry, analysis and algebra.

Geometry is probably the most familiar to us. It begins with a sense of space, codified into principles that describe how static things in space relate to each other, like a triangle's sides.

When you have to consider things that move and change with time, you come to analysis, a field that includes calculus, whether it's integral or differential calculus, or its many variations.

Algebra is what allows us to process knowledge in terms of numbers, symbols and equations – and it is the backbone of formal higher mathematics. Algebra encompasses such esoterica as group theory (the study of groups, where groups are sets of elements that satisfy certain properties), graph theory (which studies how things are interconnected) and topology (the mathematics of shapes that can be deformed continuously, without breaking and reattaching).

ANIMAL INSTINCTS

The debate over whether our sense of exact numbers is innate has often turned to animals for support. If our distant cousins can be shown to share certain mathematical abilities, then that implies our own must predate the development of culture. Certainly, some individual animals have been shown to display remarkable talent. Alex, an African grey parrot trained by Irene Pepperberg, could correctly identify sets of between two and six objects 80 per cent of the time. Ai, a chimpanzee trained by Japanese primatologist Tetsuro Matsuzawa, could do much the same.

But too much emphasis is placed on research involving animals, says Rafael Núñez at the University of California, San Diego, at the expense of data from human cultures that have sophisticated languages and yet don't show exact numerosity (see main story).

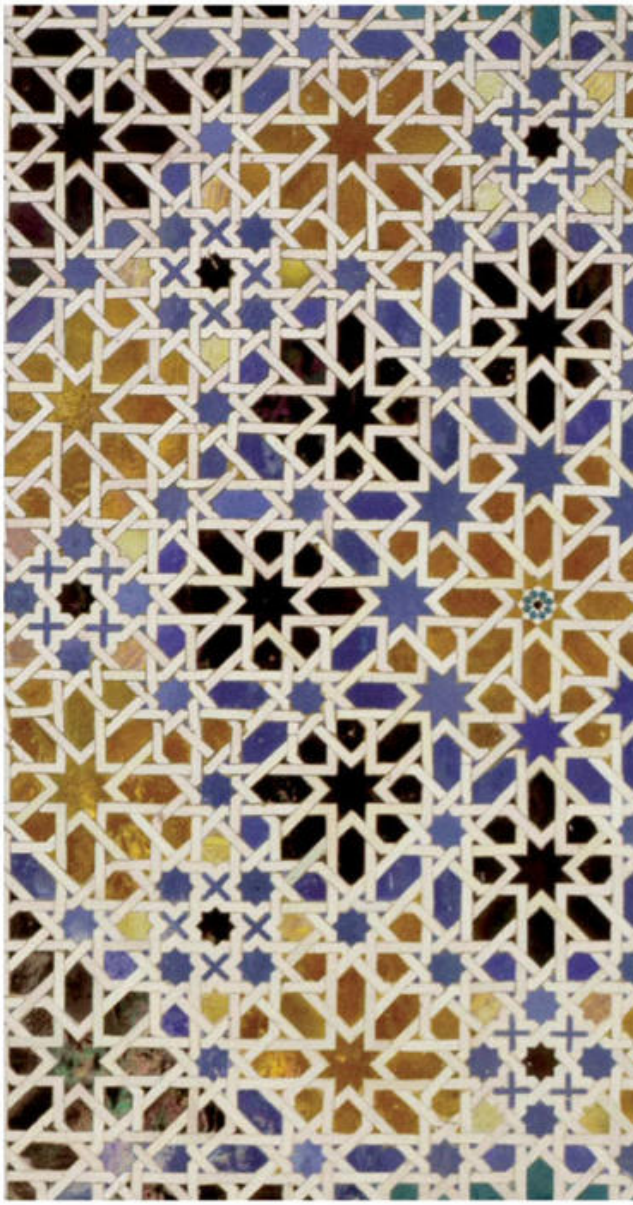
The animals aren't grasping the symbolic meaning of numbers, he argues. Instead they are simply learning about numbers by association after thousands of tests. It's not unlike how we train animals to do all sorts of things they wouldn't do in the wild. "Are elephants capable of standing on one leg on a little stool wearing a funny hat? Well, yes, if you train them for a long time," says Núñez.

But there is growing evidence that animals are capable of feats approaching numerical ability in their natural habitats. In the early 1990s, lions were shown to distinguish between recordings of one lion and three lions roaring. In 2017, at a meeting of the Royal Society, London, researchers reported that some frogs can listen to the calls of competing frogs and either match these calls in number or go one better.

Brian Butterworth of University College London believes such findings show that animals are able to discriminate solely on the basis of numerical information. "We share this with many other creatures," he says.

But these assertions remain contentious. Not everyone agrees that such results demonstrate an animal's instinct for numerosity.

RICK FRIEDMAN/CORBIS VIA GETTY IMAGES



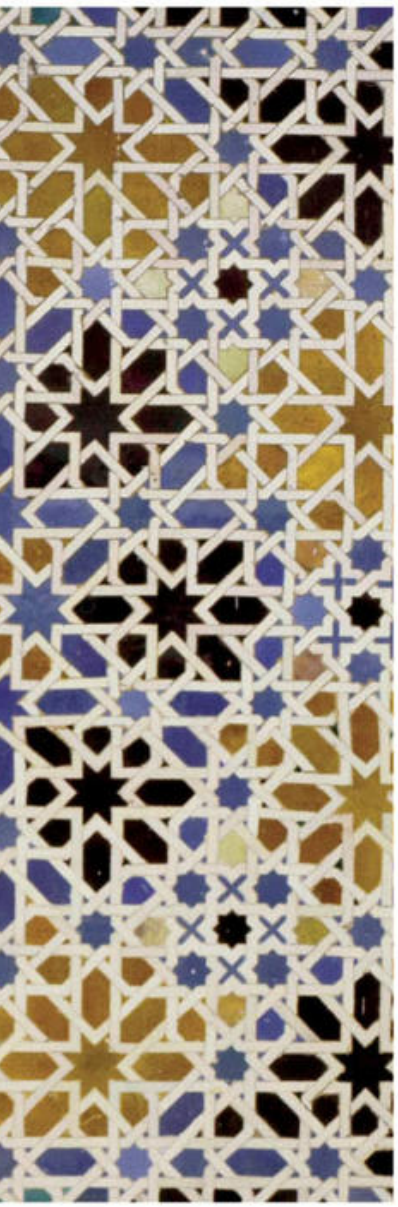
WERNER FORMAN/UNIVERSAL IMAGES GROUP/GETTY IMAGES

Some of this sophisticated mathematical language certainly develops out of our inbuilt sense for numbers or magnitudes, however imprecise it might be at birth. But it probably also leans on many other abilities: language to communicate ideas, working memory to hold and manipulate concepts, and even cognitive control to overcome the kinds of biases apparent in the gambler's fallacy.

The exact moment when culture transformed whatever instincts we may have had into a recognisable mathematical ability is unclear. One of the earliest pieces of evidence of humans dealing with numbers comes from the Border cave in the Lebombo Mountains in South Africa. There, archaeologists found 44,000-year-old bones with notches, including the fibula of a baboon etched with 29 such marks. Anthropologists think that such "tally sticks" were an aid to counting, and represent evidence for an emerging symbolic understanding related to consciously representing and manipulating numbers.

Counting and measuring hit new heights sometime around the 4th millennium BC, in the sophisticated Mesopotamian culture of the Tigris-Euphrates valley, a region in modern-day Iraq. Eleanor Robson at the University of Oxford has argued that mathematics in Mesopotamia was a cultural invention needed to keep track of days, months and years, to measure areas of land and amounts of grain, and maybe even to record weights. And as humans took to the seas, or studied the skies, we began developing the mathematics required for navigation and for tracking celestial objects. But it was always, in the beginning, a product of cultural necessity (and if you think trading-driven mathematics is a thing of the past, think again: some of the most sophisticated mathematics is being developed for trading stocks and bonds on Wall Street).

With the help of fundamental mathematical tools, humans have built an immense pyramid of mathematical knowledge (see "The pillars of mathematics", page 9). Over the past 5000 years or so, mathematics has expanded into ever more abstract domains, seemingly further removed from the processes that govern the world around us.
And yet, the more we learn about the universe's hidden workings, the more such mathematical innovation seems to describe the things we see. When David Hilbert developed a highly abstract algebra that worked in an infinite number of dimensions rather than the familiar three dimensions of space, for example, nobody could have



Mathematics helps us make sense of patterns we see in the world around us

WHY DO PEOPLE HATE MATHS?

"It is familiar to anyone writing about (or teaching) mathematics: no one very much likes the subject," writes mathematician David Berlinski in his book *One, Two, Three*.

This distaste, even fear, of mathematics is common – most of us know the feeling. Berlinski says this can be attributed to its use of arcane symbols. Symbols are strange, plus using them in the forms of theorems and proofs demands great attention, and the pay-off is never obvious (have you ever asked "how is learning algebra going to help me in real life?"). "In mathematics, something must be invested before anything is gained, and what is gained is never quite so palpable as what has been invested," writes Berlinski.

Maths anxiety, a tendency to panic when asked to perform mathematical tasks, is a very real thing. But it's incredibly difficult to study, says developmental psychologist Daniel Ansari at the University of Western Ontario in Canada. When a child displays such anxiety in school, for example, it's not clear whether it stems from an aversion to mathematical symbols, from an inability to use language to talk about mathematics or from social causes such as an overbearing parent.

foreseen its use in the emerging field of quantum mechanics. But soon after, it turned out that the state of a quantum system could best be described using such a Hilbert space – with the underlying mathematics remaining key to our attempts to make sense of the quantum world.

The ubiquity of such connections between mathematics and physics led the physicist Eugene Wigner to comment on the "unreasonable effectiveness of mathematics" at describing the natural world. To many physicists today, the success of mathematics as a language speaks to its primacy in the organisation of the universe.

Max Tegmark of the Massachusetts Institute of Technology is one of these. He thinks the universe is a mathematical structure in that it has only mathematical properties – and we are slowly uncovering this structure, brushing away the dust to reveal the theorems and proofs that underpin reality. "It used to be that it was very easy to list the small number of things in nature that you could describe with maths. Now it's very easy to list the small number of things you cannot," says Tegmark. Even biology, which long resisted mathematical rigour, is slowly succumbing: witness the proliferation of mathematics in genomics or computational neuroscience.

From this perspective, mathematics is a discovery rather than an invention. For researchers like Núñez, however, that is an overly simplistic distinction. "When the question is asked – is mathematics invented or discovered?" he says, "there is a supposition that it's exclusive. If you invented it, you don't discover it, and so on." But it is not an either-or situation, he says.

Think of *Elements*, compiled by the ancient Greek mathematician Euclid, which unified all of Greek mathematical knowledge of the time and codified the laws of geometry. Euclid based his work on a series of rules or axioms, one of the most famous being that parallel lines never meet. Over time, the patterns, regularities and relationships that emerged from these "invented" axioms were explored by other mathematicians and proved as theorems. In a sense, they were "discovering" the landscape of Euclidean geometry. But then, thousands of years later, other mathematicians decided to start with axioms that contradicted the ones Euclid set out.

Riemannian geometry, for example, which owes its name to the German mathematician Bernhard Riemann, explicitly relies on the idea that parallel lines can in fact meet. This unorthodox starting point led to the discovery of a rich vein of mathematics that Einstein would use to formulate his general theory of relativity and describe the curvature of space-time. "The world out there has all kinds of patterns and regularities and ways of behaving, and any creature that is going to build a mathematics is going to have to build it on top of regularities that are constraining the behaviour of the stuff that they encounter," says Clark.

But no matter which axioms we start off with, mathematics might not be as complete a system of thought as we like to believe. We owe that insight to Austrian logician Kurt Gödel's incompleteness theorem.

> "Some ask if mathematics is invented or discovered. It's not an either-or"

Gödel showed that within the bounds of any formal system of axioms and theorems, you can make statements that can be neither proved nor disproved. In other words, there are some questions that mathematics can ask, but it will never have the tools to answer.

In which case, perhaps it is too early for us to make any sweeping statements about mathematics being a universal truth. After all, who's to say that our little corner of the jungle is in any way representative of the whole? But physicists like Tegmark have hope. For him, the biggest hurdle to a mathematical theory of everything is a description of consciousness, the crucible of our own numerical ability. Getting maths to explain its own origins? "That's going to be the final test of the hypothesis that it's all mathematics," he says. ■

The foundations of mathematics might not be as solid as they appear, says **Richard Elwes**

It doesn't add up

IF YOU were forced to learn long division at school, you might have had cause to curse whoever invented arithmetic. A wearisome whirl of divisors and dividends, of bringing the next digit down and multiplying by the number you first thought of, it almost always went wrong somewhere. And all the while you were plagued by that subversive thought – provided you were at school when such things existed – that any sensible person would just use a calculator.

Well, here's an even more subversive thought: are the rules of arithmetic, the basic logical premises underlying things like long division, unsound? Implausible, you might think. After all, human error aside, our number system delivers pretty reliable results. Yet the closer mathematicians peer beneath the hood of arithmetic, the more they are becoming convinced that something about numbers doesn't quite add up. The motor might be still running, but some essential parts seem to be missing – and we're not sure where to find the spares.

From the 11-dimensional geometry of superstrings to the subtleties of game theory, mathematicians investigate many strange and exotic things. But the system of natural numbers – 0, 1, 2, 3, 4 and so on ad infinitum – and the arithmetical rules used to manipulate them retain an exalted status as mathematics' oldest and most fundamental tool.

Thinkers such as Euclid around 300 BC and Diophantus of Alexandria in the 3rd century AD were already probing the deeper reaches of number theory. It was not until the late 19th century, though, that the Italian Giuseppe Peano produced something like a complete set of rules for arithmetic: precise logical axioms from which the more complex behaviour of numbers can be derived. For the most part, Peano's rules seem self-evident, consisting of assertions such as if $x = y$, then $y = x$ and $x + 1 = y + 1$. It was nevertheless a historic achievement, and it unleashed a wave of interest in the logical foundations of number theory that persists to this day.

It was 1931 when a young Austrian mathematician called Kurt Gödel threw an almighty spanner in the works. He proved the existence of "undecidable" statements about numbers that could neither be proved nor disproved starting from Peano's rules. What was worse, no conceivable extension of the rules would be able to deal with all of these statements. No matter how many carefully drafted clauses you added to the rule book , undecidable statements would always be there (see "Bound not to work", page 14).

Gödel's now-notorious incompleteness theorems were a disconcerting blow. Mathematics prides itself on being the purest route to knowledge of the world around us. It formulates basic axioms and, applying the tools of uncompromising logic, uses them to deduce a succession of ever grander theorems. Yet this approach was doomed to failure when applied to the basic system of natural numbers, Gödel showed. There could be no assumption that a "true" or "false" answer exists. Instead, there was always the awkward possibility that the laws of arithmetic might not supply a definitive answer at all.

A blow though it was, at first it seemed it was not a mortal one. Although several examples of undecidable statements were unearthed in the years that followed, they were all rather technical and abstruse: fascinating to logicians, to be sure, but of seemingly little relevance to everyday arithmetic. One plus one was still equal to two; Peano's rules, though technically incomplete, were adequate for all practical purposes.

In 1977, though, Jeff Paris of the University of Manchester, UK, and Leo Harrington of the University of California, Berkeley, unearthed a statement concerning the different ways collections of numbers could be assigned a colour. It could be simply expressed in the language of arithmetic, but proving it to ›

"Gödel revealed the awkward possibility that arithmetic sometimes could not supply any answers at all"

BRETT RYDER

Bound not to work

In the 1920s, David Hilbert laid down a grand challenge to his fellow mathematicians: to produce a framework for studying arithmetic, meaning the natural numbers together with addition, subtraction, multiplication and division, with Giuseppe Peano's axioms as its backbone. Such a framework, Hilbert said, should be consistent, so it should never produce a contradiction such as 2 + 2 = 3. And it should be complete, meaning that every true statement about numbers should be provable within the framework.

Kurt Gödel's first incompleteness theorem, published in 1931, killed that aspiration dead by encoding in arithmetical terms the statement "this statement is unprovable". If the statement could be proved using arithmetical rules, then the statement itself is untrue, so the underlying framework is inconsistent. If it could not be proved, the statement is undeniably true, but that means the framework is incomplete.

In a further blow, Gödel showed that even mere consistency is too much to ask for. His second incompleteness theorem says that no consistent framework for arithmetic can ever be proved consistent under its own rules.

The *coup de grâce* was delivered a few years later, when Briton Alan Turing and American Alonzo Church independently proved that another of Hilbert's demands, that of "computability", could not be fulfilled: it turns out to be impossible to devise a general computational procedure that can determine whether any statement in number theory is true or false.

BRETT RYDER

be true for all the infinitely many possible collections of numbers and colourings turned out to be impossible starting from Peano's axioms (see "The colour of numbers", below right).

The immediate question was how far beyond Peano's rules the statement lay. The answer seemed reassuring: only a slight extension of the rule book was needed to encompass it. It was a close thing, but Gödel's chickens had once again missed the roost.

Recently, though, they seem to have found their way at least closer to home. In a book published in 2011, *Boolean Relation Theory and Incompleteness*, the logician Harvey Friedman of Ohio State University in Columbus identified an entirely new form of arithmetical incompleteness. Like Paris and Harrington's theorem, these new instances, the culmination of more than 10 years' work, involve simple statements about familiar items from number theory. Unlike Paris and Harrington's theorem, they lie completely out of sight of Peano's rule book.

Understanding what Friedman's form of incompleteness is about means delving into the world of functions. In this context, a function is any rule that takes one or a string of natural numbers as an input and gives another number as an output. If we have the numbers $x = 14$, $y = 201$ and $z = 876$ as the input, for example, the function $x + y + z + 1$ will produce the output 1092, and the function $xyz + 1$ will give 2,465,065.

These simple functions belong to a sub-class known as strictly dominating functions, meaning that their output is always bigger than their inputs. A striking fact, known as the complementation theorem, holds for all such functions. It says there is always an infinite collection of inputs that when fed into the function will produce a collection of outputs that is precisely the non-inputs. That is to say, the inputs and outputs do not overlap – they are "disjoint sets" – and can be combined to form the entire collection of natural numbers.

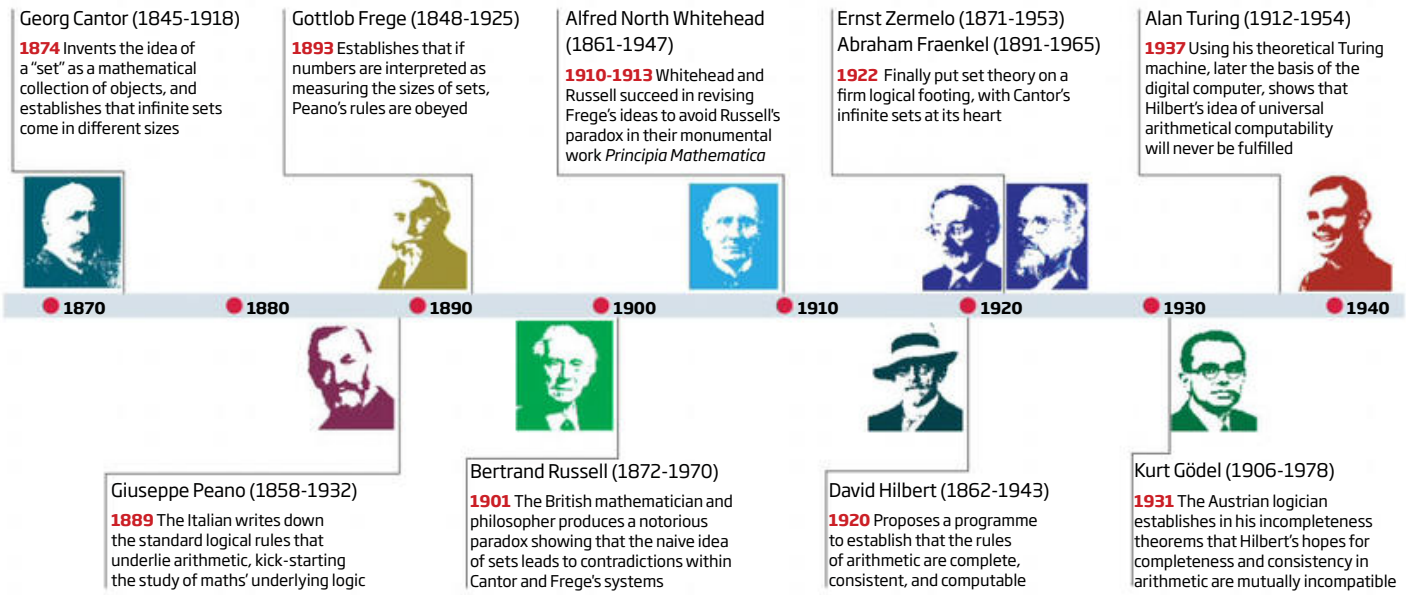
Delayed triumph

As an example, consider the basic strictly dominating function that takes a single number as its input and adds 1 to it. Here, if you take the infinite set of even numbers 0, 2, 4, 6, 8, 10… as the inputs, the corresponding outputs are the odd numbers 1, 3, 5, 7, 9, 11… Between them, these inputs and outputs cover every natural number with no overlap. The complementation theorem assures us that a configuration like this always exists for any strictly dominating function, a fact that can be deduced from Peano's rules.

Friedman's work entails adjusting the complementation theorem to pairs of a specific class of strictly dominating function known as expansive linear growth (ELG) functions.

Logical ups and downs

The quest for the logical underpinnings of arithmetic occupied many of the greatest figures in mathematics in the late 19th and early 20th centuries – but the story turned out to be infinitely more complex than first thought



Friedman identified 6561 relationships between inputs and outputs that a pair of ELG functions could exhibit in principle. For every one of these relationships, he tested the hypothesis that it would be shown by every possible pair of ELG functions.

Friedman found that Peano's rules gave a definitive yes or no answer in almost all cases. The relationship either popped up with every pair of ELG functions, or he found a specific pair whose inputs and outputs could not be linked in that way. In 12 cases, however, he drew a blank: the hypothesis could neither be proved nor disproved using Peano's axioms. What's more, it could not be proved using any reasonable extension of conventional arithmetic. With Friedman's work, it seems Gödel's delayed triumph has arrived: the final proof that if there is a universal grammar of numbers in which all facets of their behaviour can be expressed, it lies beyond our ken.

What does this mean for mathematics, and for fields such as physics that rely on the exactitude of mathematics? In the case of physics, probably not much. "Friedman's work is beautiful stuff, and it is obviously an important step to find unprovable statements that refer to concrete structures rather than to logical abstractions," says theoretical physicist Freeman Dyson of the Institute for Advanced Studies in Princeton, New Jersey. "But mathematics and physics are both open systems with many uncertainties, and I do not see the uncertainties as being the same for ›

The colour of numbers

When Jeff Paris and Leo Harrington got their glimpse of arithmetical incompleteness in 1977, they were considering a variant on a classic mathematical result called Ramsey's theorem. Suppose we have some scheme for assigning one of two colours, either red or blue, to every possible set of four natural numbers. So {1, 5, 8, 101} might be red for example, and {101, 187, 188, 189} might be blue. It is quite possible, then, that any given number will occur in some red sets and some blue sets. What Ramsey's theorem says is that, despite this, we can always find an infinite collection of numbers that is monochromatic – coloured entirely red or blue. There's nothing magic about sets of four numbers or two colours: change those to any figures you like, and the same thing works.

The theorem means order can be recovered even from highly disordered situations: even if you invent some horribly complex rule to colour your sets of numbers, you will always be able to extract an infinite monochromatic set. In theoretical computer science, for example, that permits algorithms to be constructed that allow the transfer of information through noisy channels where errors can creep in.

The variant of Ramsey's theorem considered by Paris and Harrington deals with sets of numbers that are "big", meaning that their smallest entry is less than the number of members in the set. So the set of four numbers {5, 7, 8, 100} is not deemed big as its smallest entry is 5, while the set {3, 8, 12, 100} is. If we start with a very big (but not infinite) set of natural numbers A, and again assign every set of four numbers within A either the colour red or blue, the modified version of Ramsey's theorem says we can find a monochromatic subset of A that is big. Again, the same result should hold with the numbers four and two replaced with any other numbers.

Therein lies the problem. Paris and Harrington showed that for the theorem to hold, the set A must be mind-bogglingly large – too huge, in fact, to be described by arithmetical procedures stemming only from Peano's rules.

A ladder of infinities

How big is infinity? A silly question, you might say, as infinity is infinitely big. Perhaps, but as the 19th-century German mathematician Georg Cantor proved to his contemporaries' dismay, the infinite comes in different sizes.

Take the natural numbers: 0, 1, 2, 3, 4, 5… You can go on counting these till kingdom come, so there's no doubting that the set of natural numbers is infinite. But this "countable" infinity occupies only the lowest rung of an infinite ladder. Ironically, larger infinities arise when you break down the natural numbers into subsets: the numbers 1 to 1,000,000, for example, or the odd numbers, the prime numbers, or pairs of numbers such as four and 1296.

How many such subsets are there altogether? An infinite number, of course. Cantor was able to prove that this infinity is bigger than the original countable set. This second level of infinity is the "continuum", and it is where many important mathematical objects live: the set of real numbers (the integers and all the fractional and irrational numbers that lie between them) and the complex numbers too.

And so it goes on. By looking at the collection of all possible subsets of real numbers, you find a still higher level of infinity, and so on ad infinitum. Infinity is not a single entity, but an infinite ladder of infinities, with each rung infinitely higher than the one below. Mathematicians call these different levels the "infinite cardinals".

In 1908, another German mathematician, Felix Hausdorff, conceived the idea of "large cardinals". These dwarf even the hugest of Cantor's original cardinals and are blessed with a hierarchy all their own. They are too far up even to be seen from below, and whether or not they exist is a question utterly beyond the range of all the ordinary rules of mathematics. Small wonder, then, that many mathematicians baulk at the claim that large cardinals could rescue the logical foundations of arithmetic (see main story).

"The rules we use to manipulate numbers might not be universal truths, but just our best approximation of reality"

both." The clocks won't stop or apples cease to fall just because there are questions we cannot answer about numbers.

The most severe implications are philosophical. Friedman's demonstration of incompleteness means that the rules we use to manipulate numbers cannot be assumed to represent the pure and perfect truth. Rather, they are something more akin to a scientific theory such as the "standard model" that particle physicists use to predict the workings of particles and forces: our best approximation to reality, well supported by experimental data, but at the same time manifestly incomplete and subject to continuous and possibly radical reappraisal as fresh information comes in.

Cardinal sins

That is an undoubted strike at mathematicians' self-image. Friedman's work does offer a face-saving measure, but it too is something that many mathematicians are reluctant to countenance. The only way that Friedman's undecidable statements can be tamed, and the integrity of arithmetic restored, is to expand Peano's rule book to include "large cardinals" – monstrous infinite quantities whose existence can only ever be assumed rather than logically deduced (see "A ladder of infinities", above).

Large cardinals have been studied by logicians for a century, but their intangibility means they seldom feature in mainstream mathematics. A notable exception is perhaps the most celebrated result of recent years, the proof of Fermat's last theorem by the British mathematician Andrew Wiles in 1994. This theorem states that Pythagoras's formula for determining the hypotenuse of a right angled triangle, $a^2+b^2=c^2$, does not work for any set of whole numbers a, b and c when the power is increased to 3 or any larger number.

To complete his proof, Wiles assumed the existence of a type of large cardinal known as an inaccessible cardinal, technically overstepping the bounds of conventional arithmetic. But there is a general consensus among mathematicians that this was just a convenient short cut rather than a logical necessity. With a little work, Wiles's proof should be translatable into Peano arithmetic or some slight extension of it.

Friedman's configurations, on the other hand, lay down an ultimatum: either admit large cardinals into the axioms of arithmetic, or accept that those axioms will always contain glaring holes. Friedman's own answer is unequivocal. "In the future, large cardinals will be systematically used for a wide variety of concrete mathematics in an essential, unremovable way," he says.

Not everyone is happy to take that lying down. "Friedman's work is beautiful mathematics, but pure fiction," says Doron Zeilberger of Rutgers University in Piscataway, New Jersey. He has a radically different take. The problems highlighted by Friedman and others, he says, start when they consider infinite collections of objects and realise they need ever more grotesque infinite quantities to patch the resulting logical holes. The answer, he says, is that the concept of infinity itself is wrong. "Infinite sets are a paradise of fools," he says. "Infinite mathematics is meaningless because it is abstract nonsense."

Rather than patching each hole with ever more dubious infinities, Zeilberger says we should focus our efforts on the only place where we really be sure of our footholds – strictly finite mathematics. When we do that, the incompleteness that creeps in at the infinite level will dissolve, and we can hope for a complete and consistent, albeit truncated, theory of arithmetic. "We have to kick the misleading word 'undecidable' from the mathematical lingo, since it tacitly assumes that infinity is real," he says. "We should rather replace it by the phrase 'not even wrong'. In other words, 'utter nonsense'".

Such "finitist" views are nothing new. They appeared as soon as Georg Cantor started to investigate the nature of infinity back in the late 19th century. It was a contemporary of Cantor's, Leopold Kronecker, who coined the finitist motto: "God created the integers; all else is the work of man." But most mathematicians believe we cannot dismiss infinity so easily, even if by accepting even the lowliest, most manageable form of infinity – that embodied by the "countable" set of natural numbers – we usher in a legion of undecidable statements, which in turn can only be tamed by introducing the true giants of the infinite world, the large cardinals.

The debate rages on, caught between two equally unpalatable conclusions. We can deny the existence of infinity, a quantity that pervades modern mathematics, or we must resign ourselves to the idea that there are certain things about numbers we are destined never to know. Mathematics works – we're just not sure how ■

See chapter 4 for more on the fraught relationship between infinity, mathematical logic and reality

INTRODUCING THE

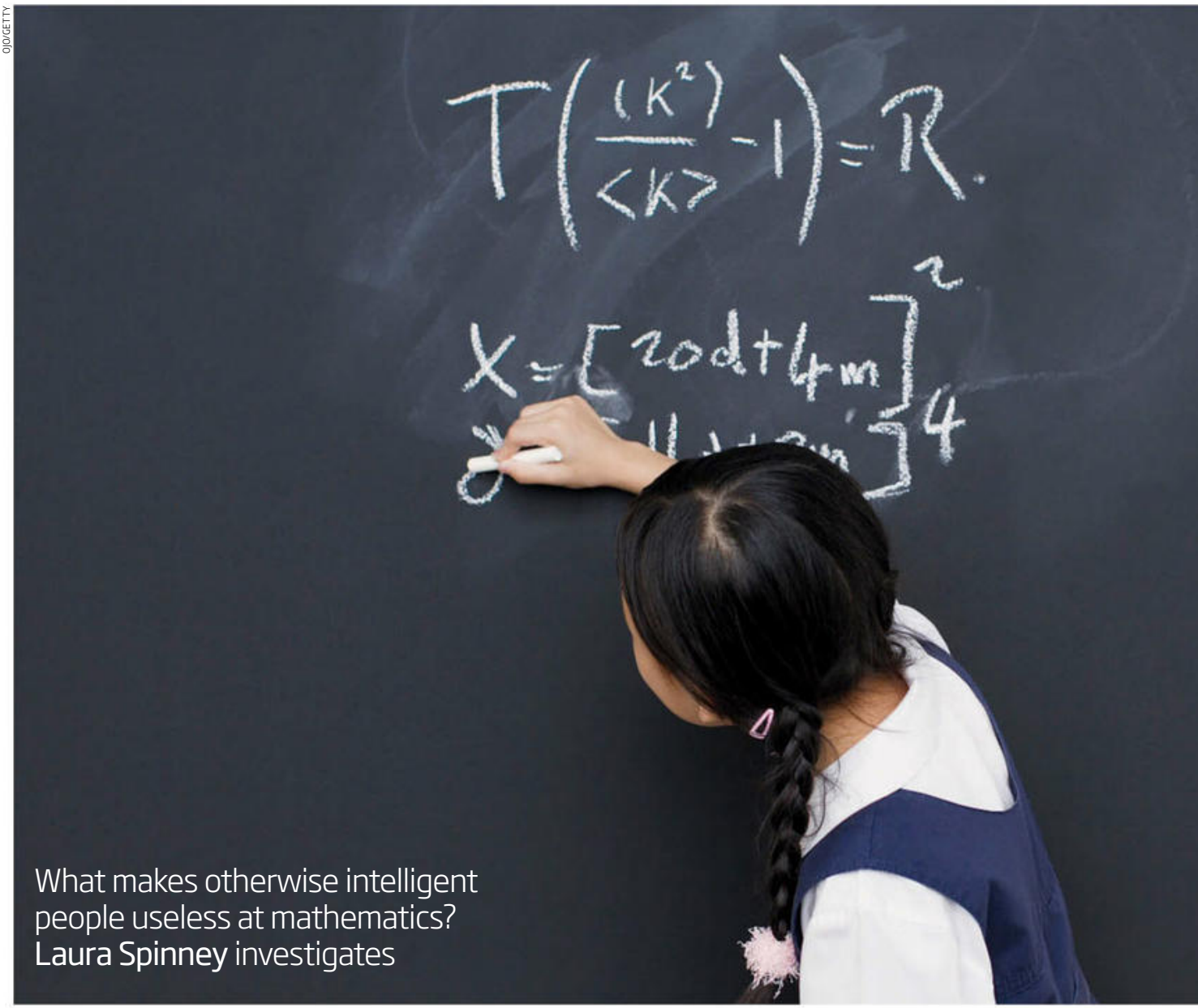
New Scientist

INSTANT EXPERT SERIES

DEFINITIVE, ENGAGING AND ACCESSIBLE GUIDES TO THE MOST IMPORTANT SUBJECTS IN SCIENCE.

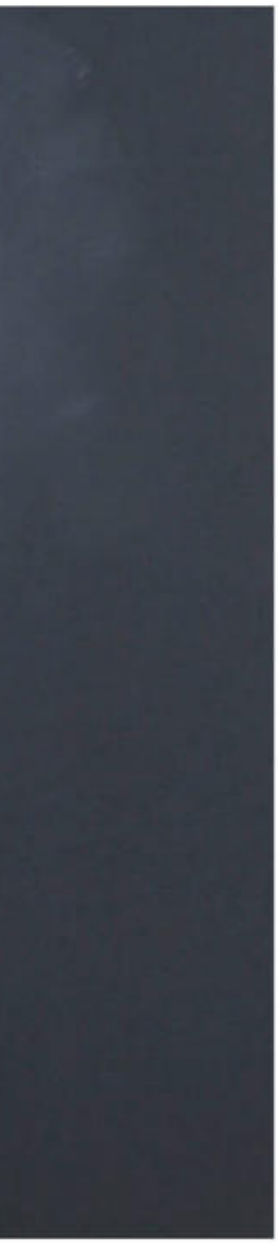
NewScientist newscientist.com/instant-expert-books

OJO/GETTY



What makes otherwise intelligent people useless at mathematics? **Laura Spinney** investigates

No good with numbers



People who struggle with arithmetic may have no problem with more conceptual maths

JILL, 19, from Michigan, wanted to go to university to read political science. There was just one problem: she kept failing the mathematics requirement. "I was an exceptional student in all other subjects, so my consistent failure at math made me feel very stupid," she says. In fact, she stopped going to her college mathematics class after a while because, she says, "I couldn't take the daily reminder of what an idiot I was."

Jill got herself screened for learning disabilities. She found that while her IQ was above average, her numerical ability was equivalent to that of an 11-year-old. The diagnosis came partly as a relief, because it explained a lot of difficulties she had in her day-to-day life. She can't easily read a traditional, analogue clock, for example, and always arrives 20 minutes early for fear of being late. When it comes to paying in shops or restaurants, she hands her wallet to a friend and asks them to do the calculation, knowing that she is likely to get it wrong.

"Dyscalculics fail to see the connection between a set of objects and the numerical symbol that represents it"

Welcome to the stressful world of dyscalculia, where numbers rule because inhabitants are continually trying to avoid situations in which they have to perform even basic calculations. Despite affecting about 5 per cent of people – roughly the same proportion as are dyslexic – dyscalculia has long been neglected by science, and people with it incorrectly labelled as stupid. But researchers have begun to get to the root of the problem, bringing hope that dyscalculic children will start to get specialist help just as youngsters with dyslexia do.

For hundreds of millions of people this really matters. "We know that basic mathematical fluency is an essential prerequisite for success in life, both at the level of employment and in terms of social success," says Daniel Ansari, a cognitive neuroscientist at the University of Western Ontario in London, Canada. A report published in October 2008 by the UK government claimed that dyscalculia cuts a pupil's chances of obtaining good exam results at age 16 by a factor of 7 or more, and wipes more than £100,000 from their lifetime earnings. Early diagnosis and extra teaching could help them avoid these pitfalls.

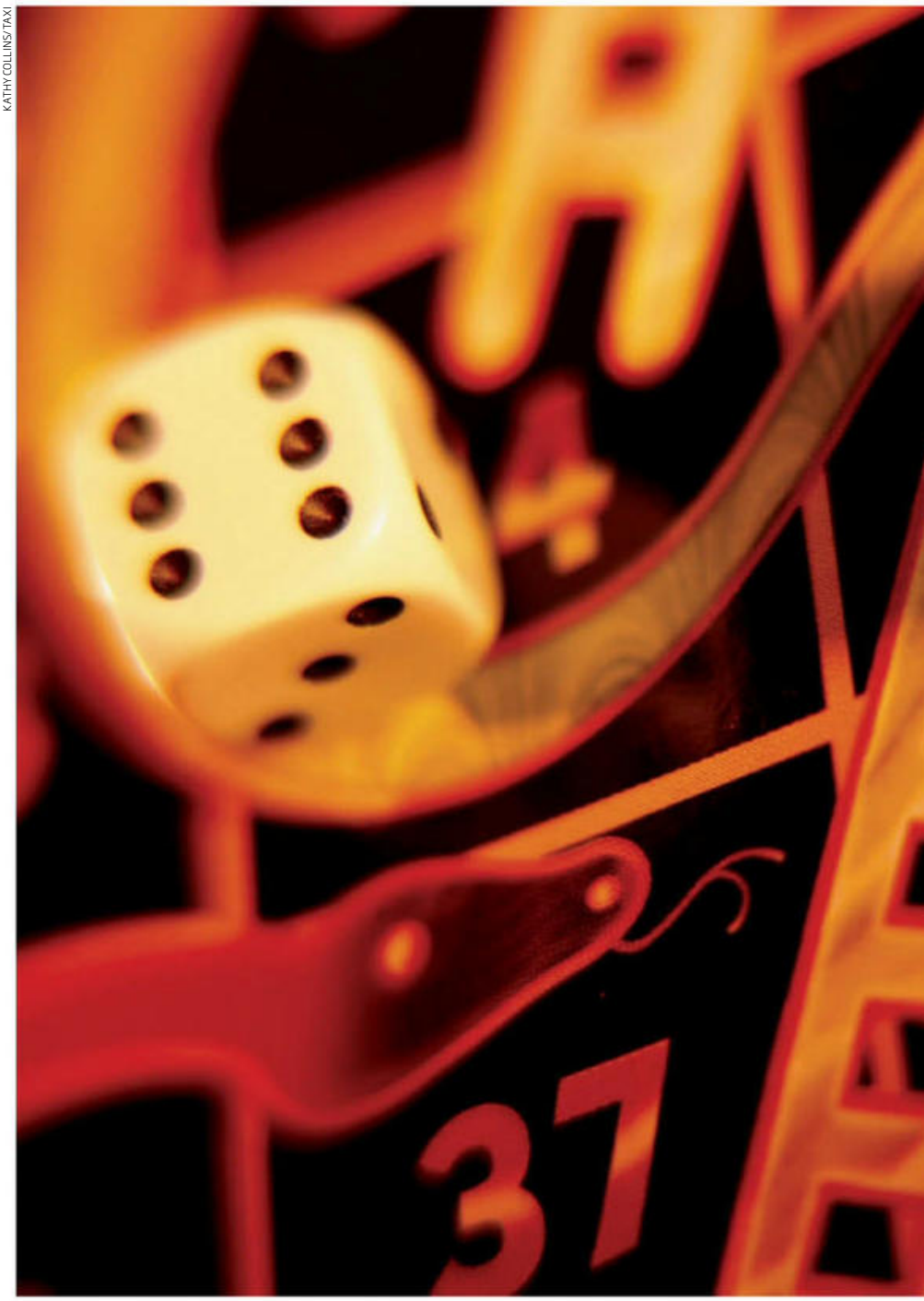
People with dyscalculia, also known as mathematics disorder, can be highly intelligent and articulate. Theirs is not a general learning problem. Instead, they have a selective deficit with numerical sets. Put simply, they fail to see the connection between a set of objects – five walnuts, say – and the numerical symbol that represents it, such as the word "five" or the numeral 5. Neither can they grasp that performing additions or subtractions entails making stepwise changes along a number line.

This concept of "exact number" is known to be unique to humans, but there is long-standing disagreement about where it comes from (see "The origin of mathematics", page 6). One school of thought argues that at least some elements of it are innate, and that babies are born with an exact-number "module" in their brain. Others say exact number is learned and that it builds upon an innate and evolutionarily ancient number system which we share with many other species. This "approximate number sense" (ANS) is what you use when you look at two heavily laden apple trees and, without actually counting the apples, make a judgement as to which has more. In this view, as children acquire speech they map number-words and then numerals onto the ANS, tuning it to respond to increasingly precise numerical symbols.

The debate over exact number is directly relevant to dyscalculics, as tackling their problem will be easier if we know what we are dealing with. If we have an innate exact number module that is somehow faulty in people with dyscalculia, they could be encouraged to put more faith in their ability to compare magnitudes using their ANS, and learn to use calculators for the rest. However, if exact number is learned, then perhaps dyscalculia could be addressed by teaching mathematics in ways that help with the process of mapping numbers onto the ANS.

So how do the two models stand up? The innate number module theory makes one obvious prediction: babies should be able to grasp exact numbers. This was explored in the early 1990s. Using dolls, a screen and the fact that babies stare for longer at things that surprise them, developmental psychologist Karen Wynn, then at the University of Arizona in Tucson, showed that five-month-old infants could discriminate between one, two and three. They look for longer if the number of dolls that come out from behind the screen does not match the number that went in.

Some teams have taken a different approach to show that we are born with a sense of exact number. They argue that if exact number is learned, it ought to be influenced by language. Back in the 2000s Brian Butterworth from University College London did tests of exact number on children aged 4 to 7 who spoke only Warlpiri or Anindilyakwa, two Australian languages that contain very few number words. He found no difference in performance between the indigenous children and a ›



KATHY COLLINS/TAXI

control group from English-speaking Melbourne. This, he says, is evidence that "you're born with a sense of exact number, and you map the counting words onto pre-existing concepts of exact numbers".

Both of these approaches, however, have been criticised. Neuroscientist Stanislas Dehaene points out that Wynn's finding also fits the rival theory – that babies enter the world with only an intuition about approximate number. This is because the ANS is concerned with ratios, so is reasonably reliable when the numbers involved are small, but falls off as the proportional size difference shrinks. A size ratio of 1:2 is more easily discernable than 9:10. Wynn tested babies on small numbers and, as Dehaene points out, "one versus two is a large ratio".

Count on learning

What is more, Dehaene has worked with an Amazonian tribe whose language only contains words for numbers up to five, and says it provides good evidence for the idea that exact number is learned (see "One, two, lots", opposite).

Supporters of the idea that exact number is learned also point to research showing how young children actually acquire an understanding of numbers. First they learn what the number word "one" means, then "two" and so on until, around the age of 4, they suddenly grasp the underlying concept of the number line and counting. "There is something very special occurring in development with exact numbers, and with the understanding of number words," says Dehaene.

For now, the idea that exact number is learned has the upper hand, suggesting that dyscalculia is a learning problem. To complicate things further, however, new research indicates that this may only be part of the story.

It was long thought that the ANS contributes little to performance in mathematics. As it is essential for survival skills such as foraging, it was assumed that everyone would have comparable abilities with approximate number. This myth was exploded in 2008 when Justin Halberda of Johns Hopkins University in Baltimore, Maryland, tested the ANS in 64 volunteers who were 14-year-olds and was "blown away" by the variability he found.

The teenagers, all of whom fell within the normal range for numeracy, watched an array of dots made up of two colours flash onto a computer screen. In each case, they had to say which colour was more numerous.

"First children learn what 'one' means, then 'two', and so on until they suddenly grasp the underlying concept"

As expected, their judgements became less accurate as the difference in size between the two sets shrank to nothing. The surprise was how much faster accuracy fell off in some kids than in others. The poorest performers had difficulty with size ratios as large as 3:4.

There was a further surprise in store when the team compared the teenagers' ANS scores with their mathematics test results from the age of 5 and up. "I literally jumped out of my seat when I saw the correlation going all the way back to kindergarten," says Halberda. The link remained even after IQ, working memory and other factors had been controlled for, and

ONE, TWO, LOTS

Amazonian hunter-gatherers called the Mundurucú only have words for numbers up to 5. Does this affect the way they think about mathematical problems? Experts who think that the human concept of exact number is innate would predict not. However, Stan Dehaene of the Collège de France in Paris is among a growing number who believe that exact number is learned and therefore affected by our culture. He decided to test this idea with the Mundurucú.

Working with his colleague in the field, Pierre Pica, and others, Dehaene has found that the Mundurucú can add and subtract with numbers under 5, and do approximate magnitude comparisons as successfully as a control group. But last year the team discovered a big cultural difference. They asked volunteers to look at a horizontal line on a computer screen that had one dot at the far left and 10 dots to the right. They were then presented with a series of quantities between 1 and 10, in different sensory modalities – a picture of dots, say, or a series of audible tones – and asked to point to the place on the line where they thought that quantity belonged.

English-speakers will typically place 5 about halfway between 1 and 10. But the Mundurucú put 3 in the middle, and 5 nearer to 10. Dehaene reckons this is because they think in terms of ratios – logarithmically – rather than in terms of a number line. By the Mundurucú way of thinking, 10 is only twice as big as 5, but 5 is five times as big as 1, so 5 is judged to be closer to 10 than to 1.

The team conclude that "the concept of a linear number line appears to be a cultural invention that fails to develop in the absence of formal education". With only limited tools for counting, the Mundurucú fall back on the default mode of thinking about number, the so-called "approximate number system" (ANS). This is logarithmic, says Dehaene. When it comes to negotiating the natural world – sizing up an enemy troop or a food haul – ratios or percentages are what count. "I don't know of any survival situation where you need to know the difference between 37 and 38," he says. "What you need to know is 37 plus-or-minus 20 per cent."

it only held for mathematics, not for other subjects. A subsequent larger study, including some children with dyscalculia, confirmed the suspicion that those with the number disorder had markedly lower ANS scores than children with average ability. This implicates a faulty ANS in dyscalculia.

Case closed? Not quite. The problem is that two other groups have come up with conflicting findings. In 2007, Laurence Rousselle and Marie-Pascale Noël of the Catholic University of Louvain (UCL) in Belgium reported that dyscalculic children, when asked to compare the magnitude of collections of sticks – say, five sticks versus seven – performed no worse than controls. However, they struggled when asked to circle the larger of two numerals, such as 5 and 7. Ansari's team has obtained a similar result. Both teams conclude that in dyscalculic children the ANS works normally, and the problem comes in mapping numerical symbols onto it.

How to account for these contradictory findings? Halberda, Ansari and Dehaene believe that there may be different types of dyscalculia, reflecting different underlying brain abnormalities. So in some dyscalculic individuals, the ANS itself is damaged, while in others it is intact but inaccessible so that individuals have problems when it comes to mapping number words and numerals onto the innate number system.

The existence of such subtypes would make dyscalculia harder to pin down, and make it difficult to design a screening programme for schoolchildren. At the moment, the condition goes widely unrecognised, and testing is far from routine. But where it is tested for, the tests are relatively crude, relying on the discrepancy between the child's IQ or general cognitive abilities and their scores in mathematics. Nevertheless, perhaps one day all children entering school will be assessed for various types of dyscalculia.

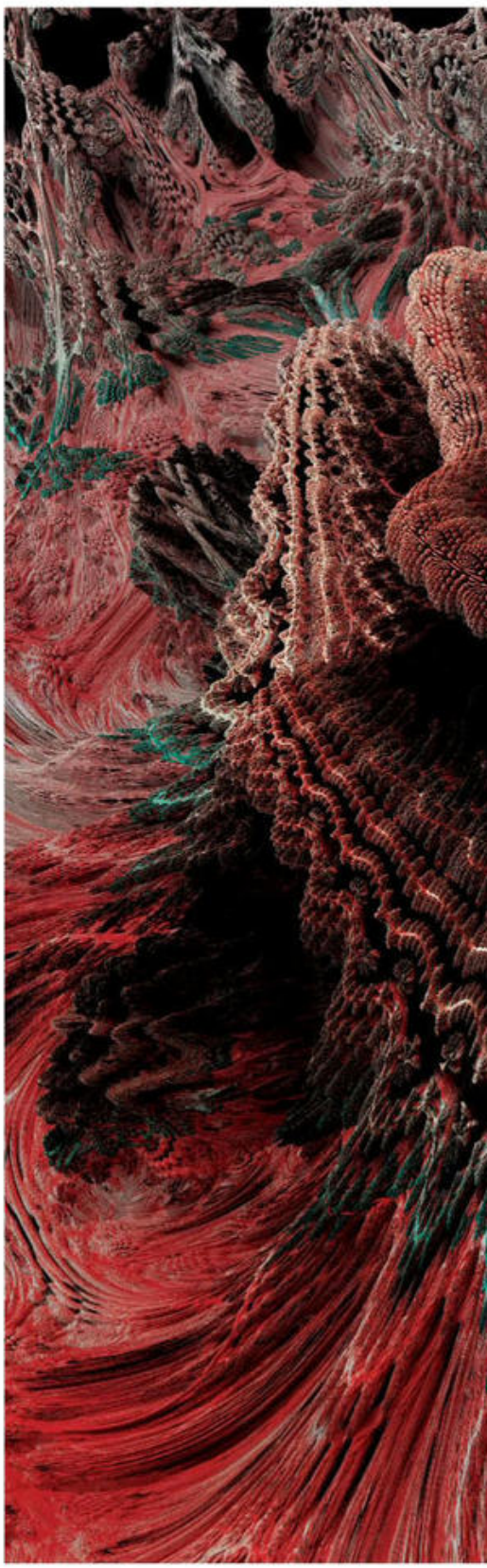
Even those researchers who remain convinced that dyscalculia is caused by a faulty exact number module believe that intervention could help. "After all, genetics isn't destiny – well, not entirely – and the brain is plastic," says Butterworth. But there is no panacea, he fears. "It may be the case that the best we can do is teach them strategies for calculation, including intelligent use of calculators, and get them onto doing more accessible branches of mathematics, such as geometry and topology."

Ansari also points out that children with dyscalculia could be helped immediately by practical measures already in place in schools for pupils with dyslexia, such as extra time in exams. And, of course, simply recognising dyscalculia as a problem on a par with dyslexia would make a huge difference. As Jill says, now that she knows what her problem is, "it's easier to have the confidence and the perseverance to keep working until I get it". That, in turn, means the condition becomes less damaging to her self-esteem and perhaps, ultimately, to her chances in life. ■

Laura Spinney is a writer based in Lausanne, Switzerland

Seduced by numbers

Mathematician **Manya Raman Sundström** thinks some of us have an inbuilt maths drive, a bit like a sex drive



MATHEMATICIANS are famous for the lengths they go to when solving problems. To crack Fermat's Last Theorem, Andrew Wiles worked in isolation for more than six years. And Thomas Hales produced a body of work consisting of 250 pages of notes and 3 gigaytes of computer programs to solve Kepler's Conjecture, a problem open since 1611 regarding the most efficient way to stack cannonballs.

What is it that motivates mathematicians to go to these extremes? It seems there is something compelling, almost seductive, about their subject. Could there be some sort of drive, similar to the sex drive? In other words, something that we could call a "maths drive" that urges us to find new mathematical explanations and truths?

As strange as this idea sounds, it is not without precedent. In 2000, the psychologist Alison Gopnik suggested, in full seriousness, that finding an explanation is like having an orgasm. Similarly, the physicist Robert Oppenheimer, a father of the atomic bomb, claimed that "understanding is a lot like sex. It's got a practical purpose, but that's not why people do it normally."

Can intellectual pursuit be as compelling as bodily urges? It might be going too far to claim that the drive to do mathematics has evolutionary roots, but perhaps not too far to suggest that it could be as rooted as the desire to reproduce – and that the production of meaningful, significant mathematics might be just as satisfying as sex.

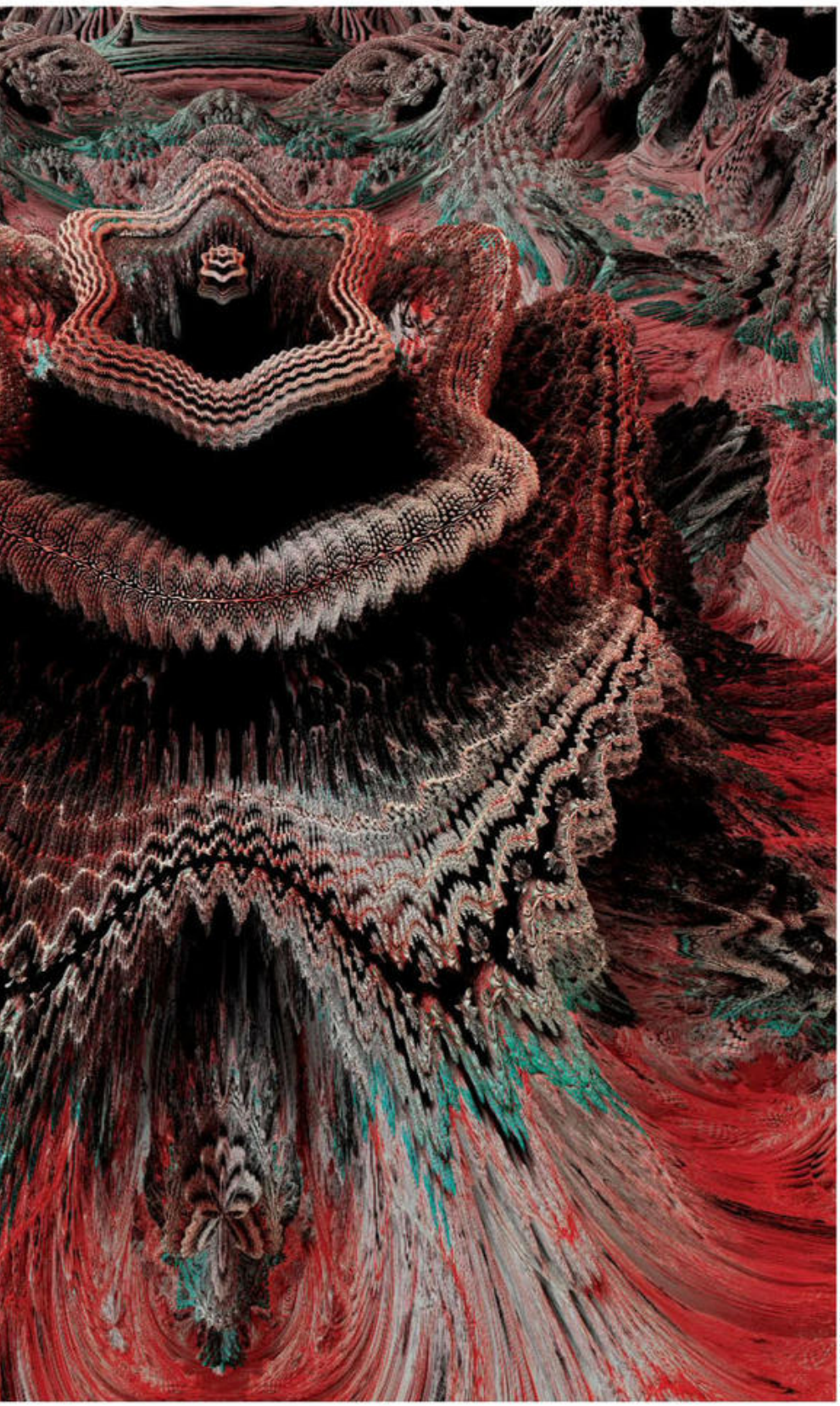
At the core of this hypothesis is a claim that doing mathematics is, at least in part, aesthetic. It is a human trait to hunt for what is beautiful, and we do so because beauty is compelling. I contend that the same is true of mathematics. Beauty – or aesthetics more generally – is not just a by-product of the subject. It isn't that you look back at the end of day and notice that a proof or definition is beautiful. It seems to be that beauty is an essential part of the process. In her article "The role of the aesthetic in mathematical inquiry", Nathalie Sinclair of Simon Fraser University, in Canada, finds that aesthetic sensibilities help guide the mind and maintain interest in a problem, as well as influencing the choice of problems to work on and the quest to find solutions to them.

This is not to say that all mathematical work is beautiful. Some proofs are tedious and long. Some, like Hales's proof of Kepler's conjecture, require computer code and are difficult to check. And it is not clear that aesthetic experiences are uniform. What is beautiful to a geometrician might not be to an algebraist. What was beautiful to you as a graduate student might not be after 20 years of research.

"Can intellectual pursuit be as compelling as bodily urges?"

Although research on the nature of mathematical beauty is under way in several fields – such as philosophy, psychology and education – there are still many open questions. What do we mean by beauty? Is it objective or subjective? Can equations be beautiful in their own right, or must they be connected to some sort of visual or sensory representation? And how does the feeling of beauty manifest itself in the brain?

Answers are beginning to emerge. For example, a study led by Semir Zeki at University College London published in 2014 involved scanning the brains of



LAGUNA DESIGN/GETTY

Art can be made from equations, but equations can be works of art in themselves

mathematicians while they viewed different formulae, such as Euler's identity, $e^{i\pi} + 1 = 0$, an equation rated as beautiful by the participants. The scans showed that the experience of mathematical beauty excited the same area of the brain as music or art.

My own research has shown that there is some consensus about what kinds of mathematical proofs are deemed beautiful. Those found to be beautiful seem to give a more immediate sense of why the claim is true. For instance, a geometric proof of the relationship between the sides of a right-angled triangle, which compared areas of small triangles inside it, was considered more aesthetically pleasing than an algebraic proof. This is probably because the algebraic proof gives no immediate sense of why the theorem is true.

Whatever we mean by the term "mathematical beauty" and how we judge it, there is no doubt that aesthetics plays a significant role in the working life of mathematicians. In 2014, after she won the Fields Medal, the maths world's Nobel prize, the late mathematician Maryam Mirzakhani talked about "the beauty of math" that one can appreciate after a lot of hard work. But how many children work through their years of schooling without experiencing this kind of appreciation? If there really is a "maths drive", at least in some proportion of the population, do we do enough to tap it?

It is not obvious whether the beauty of mathematics can be conveyed at the school level, but this question is not one that has received a great deal of attention. School lessons tend to be centred on a standard set of mathematical topics and processes. There has been little discussion of aesthetics, despite its motivational capacity.

In this, it seems we are failing to convey the true nature of mathematics. Teaching maths solely in terms of procedures such as practising sums is like teaching music through practising scales without ever exposing children to Beethoven.

When experiencing a moment of true mathematical understanding – grasping why something is so, or seeing how everything hangs together – you can feel a sense of meaningfulness, connection and purposefulness, just as you might with poetry or music. Perhaps this was what the prolific mathematician Paul Erdős meant when he claimed that certain proofs were so perfect they were divine. ■

ELLEN PORTEUS

Pairing the primes

What makes prime numbers clump in twos? If only we knew, says mathematician **Vicky Neale**

IT WAS the British mathematician G.H. Hardy who popularised the idea that youthful brains do the best maths. "I do not know of a major mathematical advance initiated by a man past fifty", he wrote in *A Mathematician's Apology*, a lament for the decline of his own creativity that he published in 1940 at the age of 62. "If a man of mature age loses interest in and abandons mathematics, the loss is not likely to be very serious either for mathematics or for himself."

If blooming youth is the rule, Yitang Zhang is a definite exception. For the best part of a decade after completing his PhD, he wasn't even working as a mathematician, instead doing odd accounting jobs around Kentucky. At one point he did a stint working in a Subway fast-food restaurant. When he announced a mathematical breakthrough that had eluded his peers for a couple of centuries, he was 57.

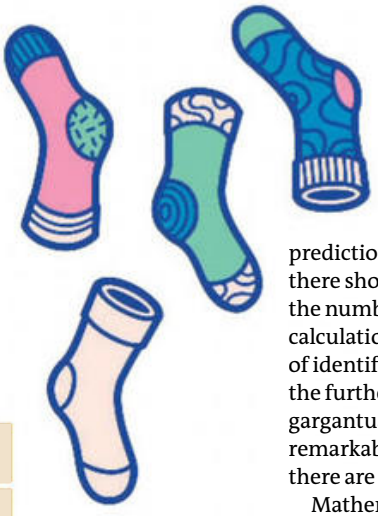
What Zhang made public in 2013 wasn't a proof of the hallowed "twin primes conjecture", but it was a significant step towards one. And even if things haven't quite panned out in the years since, he has inspired work that is promising new insights into the prime numbers, the most beguiling numbers of all.

Primes are those numbers greater than 1 that are divisible only by 1 and themselves. The sequence begins 2, 3, 5, 7, 11, 13, 17, 19 and goes on… well, as long as you like. Primes underpin modern cryptography, keeping your credit card details safe when you shop online. But their true power lies in the crucial role they play in number theory, the branch of mathematics concerned with the properties of whole numbers.

Primes are the fundamental entities from which we make all numbers, because any number that is not prime can be obtained by multiplying other primes together. "It's the same idea as in chemistry, where you might try to understand some complicated compound by understanding the atomic elements which it is made from and how they are joined together," says James Maynard, a mathematician at the University of Oxford.

The fascination with primes goes back at least as far as the ancient Greeks. In *The Elements*, Euclid came up with a beautiful proof that there are infinitely many primes, so there is no largest prime number.

Let's assume for a moment you have a list of all the prime numbers. Multiply all these together, then add 1, and you get a number that, by definition, cannot be divided exactly by any of the primes used to make it: 1 will always be left over as a remainder. Either it is divisible by another prime not on the list, or it is itself prime – so the original list must be incomplete. You can repeat this reasoning with any initial list of primes, so it follows ›



Sift it out

The sieve of Eratosthenes offers a simple way to find all the primes up to any given number. Cross out 1, which is by definition not a prime. Then cross out all multiples of 2, 3, and progressively multiples of any number not yet crossed out, for example 5 and 7. What you're left with are the primes (circled)

	(2)	(3)	4	(5)	6	(7)	8	9	10
(11)	12	(13)	14	15	16	(17)	18	(19)	20
21	22	(23)	24	25	26	27	28	(29)	30
(31)	32	33	34	35	36	(37)	38	39	40
(41)	42	(43)	44	45	46	(47)	48	49	50
51	52	(53)	54	55	56	57	58	(59)	60
(61)	62	63	64	65	66	(67)	68	69	70
(71)	72	(73)	74	75	76	77	78	(79)	80
81	82	(83)	84	85	86	87	88	(89)	90
91	92	93	94	95	96	(97)	98	99	100

that no finite list of primes contains them all.

That's crystal clear – but when it comes to theorems about patterns governing prime numbers, and in particular where they fall on the number line formed by placing all whole numbers in order, things rapidly become distinctly fuzzy.

At the end of the 19th century, Frenchman Jacques Hadamard and Belgian Charles de la Vallée Poussin independently proved what's known as the prime number theorem, which gives an estimate of the number of primes that are smaller than a million, or a trillion, or indeed any value. This theorem tells us that, on average, the primes get more spread out as we go along the number line. That fits neatly with our experience of the primes up to 100, say: the first few, 2, 3 and 5, are squashed up close, whereas there's a big gap between the two biggest primes less than 100: 89 and 97.

But right after 100 we find the primes 101, 103, 107 and 109 all bunched up together. While it is true that bigger primes get more spread out, that's just on average: look closely, and their behaviour is more nuanced.

And that's where the twin primes conjecture comes in. Apart from 2 and 3, there can't be any pairs of consecutive numbers that are both prime – one would have to be an even number, divisible by 2. But as those first primes after 100 suggest, there are many pairs of primes that differ by 2, such as 3 and 5, or 41 and 43, or 107 and 109.

Infinite twins

The twin primes conjecture predicts that, just as there are infinitely many primes, there are infinitely many pairs of these twin primes: our supply will never run out. There are good reasons to think this is the case. The first is that, with the help of computers, we have found many large twin primes. It might be, though, that the computer has found the largest there is. More compellingly, mathematicians have a model to make predictions about how many twin primes there should be up to a given point along the number line. When checked against calculations made by a computer capable of identifying twin primes out into the furthest reaches, where the truly gargantuan numbers live, the model is remarkably accurate – and it predicts there are infinitely many twin primes.

Mathematicians, though, need absolute certainty, a rigorously reasoned argument that leaves no room for doubt, as with Euclid's "proof by contradiction" argument that there are infinitely many primes. Yet even after grappling with the twin primes conjecture for hundreds of years, mathematicians have so far failed to come up with such a proof.

Hence the shock in 2013, when Zhang proved that there are infinitely many pairs of consecutive primes with a gap less than 70 million. By this point Zhang was a lecturer at the University of New Hampshire, but he had published next to nothing, so there was no suggestion that something like this was in the offing. His watertight proof made him a mathematics superstar overnight. He was inundated with job offers from prestigious institutions such as the University of California, Santa Barbara, where he now works.

Even more remarkable was that Zhang's breakthrough exploited an approach that most of the best mathematical minds had ruled out. This "sieve method" started with the ancient Greek mathematician Eratosthenes, who used it as a handy way to shake out prime numbers from the rest. In the case of finding all the primes up to 100, say, it relies on methodically crossing out all the numbers that are not prime (see "Sift it out", left). But that is too blunt an instrument to locate particular patterns of primes, so mathematicians have refined their sieving tools in various ways over the centuries.

Just over a decade ago, Daniel Goldston, János Pintz and Cem Yıldırım came up with a modified version of the sieve that came tantalisingly close to proving there are infinitely many pairs of primes that differ by at most 16. To make it work, however, they had to assume another unproven conjecture was true. This is a well-established way to make progress, but means the result doesn't amount to a complete proof. Zhang, on the other hand, was able to modify the sieve method so as not to rely on unproven assumptions.

Proving there are infinitely many consecutive primes separated by at most 70 million might sound distinctly

Prime problems

The riddle of the never-ending pairs is not the only mystery of the prime numbers

Goldbach's conjecture

This is the prediction that every even number above 4 can be written as a sum of two odd prime numbers – for example, 10 = 3 + 7, and 78 = 31 + 47. Proposed by Christian Goldbach in 1742, it remains unproven.

Infinite Germains

A Germain prime, named after Sophie Germain, is one that gives another prime if you double it and add 1. For example, 29 is prime, and (29 x 2) + 1 = 59 is also prime, so 29 is a Germain prime. Mathematicians expect that there are infinitely many Germain primes, but no one can prove it.

The Riemann hypothesis

In 1859, Bernhard Riemann put forward an idea about where the Riemann zeta function takes the value zero. Proving this conjecture would reveal more about the distribution of the primes. It is one of the Clay Mathematics Institute's seven Millennium Problems – prove Riemann's idea and you win $1 million.

unimpressive when the goal is 2, but 70 million is a lot less than infinity. What's more, this was the first time anyone had managed to prove there are infinitely many primes with a gap less than some fixed finite number. "Just to have a number was extraordinary," says Andrew Granville, a number theorist at the University of Montreal and University College London. "Everybody had tried to find a proof along these lines and I really didn't think it was possible."

As soon as the proof was published, mathematicians scrambled to understand Zhang's approach. The limit of 70 million was not the best that his argument would give, so others set about tightening up the details of the proof. The charge was led by Scott Morrison of the Australian National University, and subsequently Fields medallist

> "Zhang's watertight proof made him a mathematics superstar overnight"

Terry Tao of the University of California, Los Angeles, who started an online Polymath collaboration to tackle the problem more systematically. The idea with Polymath projects is that all contributors can work on an unsolved problem, collaborating entirely in public on blogs and wikis.

It worked beautifully in this case: within months the collaboration was able to prove that there are infinitely many pairs of primes where the gap is less than or equal to 4680. But then progress dried up. The Polymathematicians had squeezed the best they could out of Zhang's argument, and needed new tools to go further.

It took a fresh perspective from Maynard, then a postdoc at the University of Montreal in Canada, to make the gap shrink again. Revisiting the work of Goldston, Pintz and Yıldırım, he found a new way to use a sieve that was both simpler than Zhang's and gave a better result: there are infinitely many pairs of primes that differ by at most 600.

By April 2014, the Polymath project was back in the game and, using the new method, brought the gap down from 600 to less than or equal to 246. That is a huge improvement on 70 million, never mind infinity. And that, for now, is the state of the art: all the methods that got us this far have come up against the mathematical equivalent of a brick wall.

The trouble lies in the definition of a prime number, and the way sieve theory works. A prime number always has just one prime factor, namely itself. Sieve theory struggles when it's only looking for numbers with an odd number of prime factors. "It is sort of like a radar that's trying to scan for prime numbers but it gets lots of false positives," says Maynard. "You can't tell which bleeps come from primes and which come from numbers that look like primes but actually have two or four prime factors." This is what mathematicians call the parity problem, and right now there seems to be no way around it.

But Maynard has a sniff of something promising: a recent breakthrough, which gives a way of zooming in from the average behaviour of numbers across long intervals of the number line to work out patterns over shorter intervals. That was long thought to be incredibly hard, if not impossible, but in 2015, Kaisa Matomäki of the University of Turku in Finland and Maksym Radziwiłł, now at McGill University in Montreal, Canada, were able to do exactly that. "They showed that almost all the time, if you just pick some zoomed-in place, you'll get numbers with an even number of prime factors and numbers with an odd number of prime factors," says Maynard. "It's a technical result that is very exciting for us, because these nuts and bolts can often be applied in other areas."

Indeed, Tao has already used the insight to solve Chowla's conjecture, a "baby version" of the twin primes conjecture, which was created as a sort of stepping stone towards that proof. He looked at the sequence of numbers starting with 1×3, 2×4, 3×5, 4×6, 5×7, and showed that a number in this sequence is equally likely to have an odd number or an even number of prime factors.

Neither of these developments directly deals with the twin primes conjecture and, although Granville was "shocked" to see Matomäki and Radziwiłł's result, he is yet to be convinced it will help with the twin primes conjecture. "It's not at all clear how this will play out," he says.

Such is the nature of maths: you never quite know when painstakingly slow progress behind the scenes will suddenly fall into place for a big breakthrough. For Maynard, however, the signs are at least now hopeful. "The mere fact that people have handled the parity problem in contexts that are not too far away from the twin primes conjecture makes me optimistic". The mysteries of the twins could soon be up for grabs, but it might take another left-field hero like Zhang to make the breakthrough. ■

Wonders of numberland

Prime numbers are just the beginning of the number story. Numbers and patterns of numbers have all sort of uses both practical and impractical – even when they are entirely imaginary...

i

The imaginary number

THE rules of mathematics say that two positive numbers multiply to give a positive, and two negative numbers also multiply to give a positive. So what number could you multiply by itself to give -1? This is not a trick question – it's just that the answer is imaginary.

The square roots of negative numbers were first called "imaginary" by René Descartes in 1637. But it wasn't until the 18th century that they came to be represented as multiples of *i*, the square root of -1.

Imaginary numbers don't fit on the regular number line, so they are put on a second, independent line, with the two intersecting at zero.

The lines can be treated as axes, making imaginary numbers handy for representing things that change in two dimensions. They are regularly used to describe wave functions in quantum mechanics and to define alternating current.

Conjuring an entirely different family of numbers from thin air might seem unjustifiable. But the truth is that "real" and "imaginary" numbers are both abstract concepts. We might be more familiar with 5 than 5*i*, but neither exists in the real world.

That gives mathematicians a certain creative licence. In 1843, the Irish mathematician William Hamilton invented numbers called quaternions, using additional solutions for the square root of -1 that he called *j* and *k*. These form the basis of additional number lines that are used to construct axes capable of encoding rotations in 3D. Computer game design is one area where they have proved useful.

If you follow the same mathematical logic, then there is no reason to stop there. The octonions include seven dimensions of imaginary numbers, and the rarely used sedenions give the option of extending the total to 15. Down here, it's a world of pure imagination. *Gilead Amit*

BENFORD'S LAW

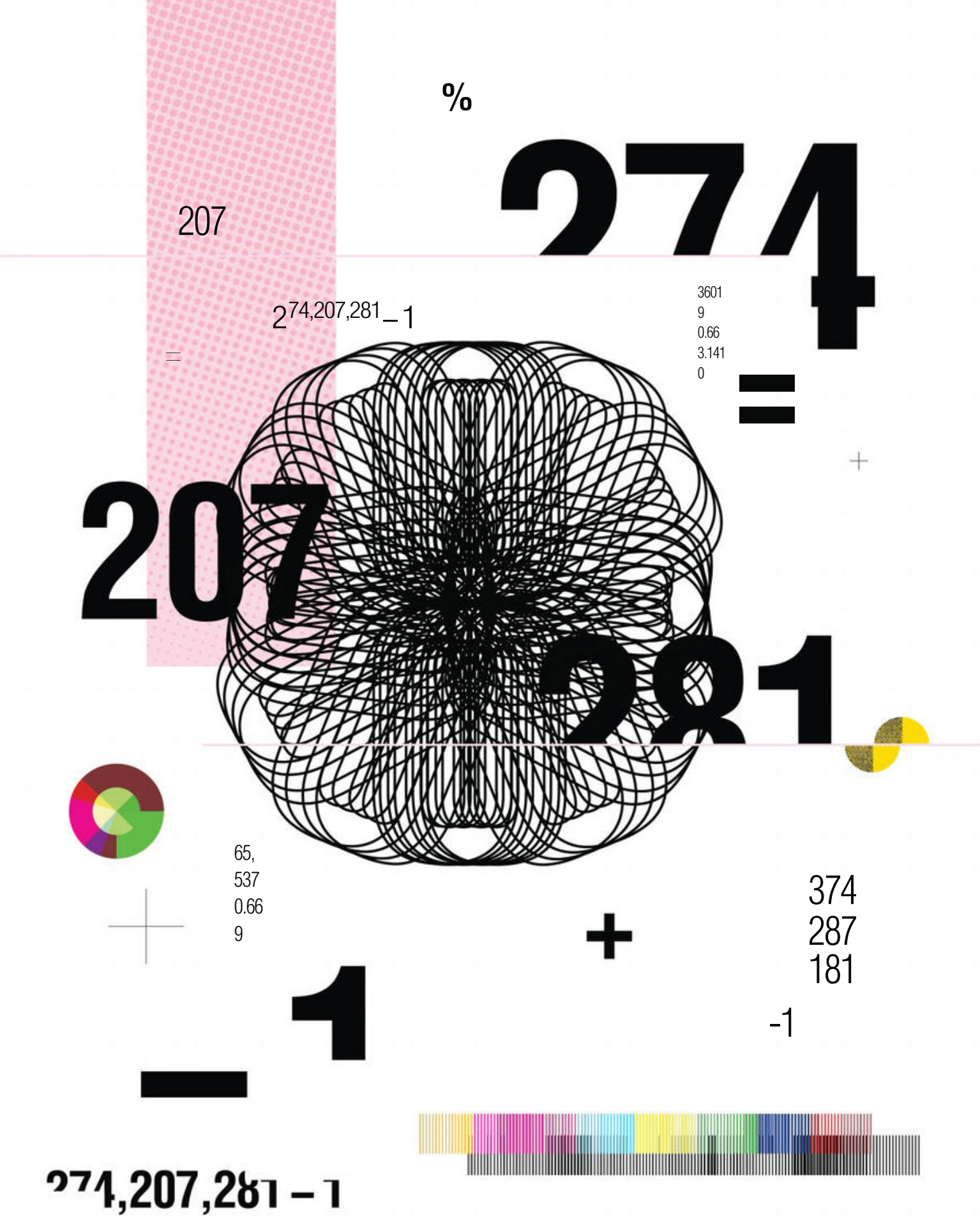
What do absent aliens, dodgy dictators and financial fraudsters have in common? Benford's law can help hunt them down.

Benford's law states that lists of numbers related to some natural or human activities will contain a particular distribution of digits. If you take a list of the areas of river basins, say, or the figures in a firm's accounts, there will always be more numbers that start with 1 than any other digit. Numbers starting with 2 are the next most common, then 3 and so on. A number will start with a 9 only 4.6 per cent of the time.

Why on earth should Benford's law exist? Drill down to the root cause of most natural processes and they depend on random things, like the jostling of atoms. That produces bell-shaped curves, where most of the values are in the middle. But if several natural phenomena are at play – which is the case in a huge number of fields – then it turns out that Benford's law is what holds.

And you can use it in lots of neat ways, especially to catch out miscreants. In 2009, for example, a suspiciously large number of vote tallies for one candidate in the Iranian elections began with a 7, suggesting vote-rigging.

The US Internal Revenue Service has scored several major successes by using Benford's law to probe firms' books for financial chicanery. And in 2013, a new application arose that brought it right back to its astronomical origins. Thomas Hair at Florida Gulf Coast University showed that the masses of thousands of confirmed and candidate exoplanets conform to the pattern. OK, it doesn't tell us where to look for ET, but it gives us confidence that our ways of seeking exoplanets aren't delivering spurious results. *Michael Brooks*


%
207
274
$2^{74,207,281}-1$
=
3601
9
0.66
3.141
0
207
281
65,
537
0.66
9
374
287
181
-1
1
274,207,281 – 1

Euler's number

Why things don't grow forever

PUT a pound in the bank. If the yearly interest were 100 per cent, then a year later you would have £2. That's simple enough. But what if instead of calculating the interest at the end of the year, the bank worked it out more regularly? It turns out this question leads us to one of the most subtle numbers in mathematics.

Say your bank paid interest twice a year but halved the rate to 50 per cent. That would take your £1 to £1.50 after 6 months, and at the end of the year you would get another 50 per cent, making £2.25 – a nice gain. If you got interest monthly but scaled down the rate accordingly, you would end up with £2.61. Do the same thing daily, reducing the interest rate in the same fashion, and you would get £2.71. The improvements get ever smaller as this process continues, and the most you could have turns out to be about £2.71828.

This number is actually a special irrational, which, like π, keeps on going forever after the decimal point. It's called Euler's number (or simply e), after the Swiss mathematician Leonhard Euler.

Euler's number doesn't just appear when computing compound interest. For instance, mix together the imaginary number i (see page 28) and e and, with a little mathematical nous, you can derive one of the most famous equations ever, Euler's identity: $e^{i\pi} + 1 = 0$. Mathematicians hold it in high regard for its beauty, cramming five of the most important numbers into a single, elegant expression.

Euler's number is also practical. It is crucial to a mathematical technique called Fourier analysis, for example, which is used by researchers who probe crystals by shining X-rays at them. Applying the analysis to the patterns that emerge helps reveal the structure of molecules such as DNA.

But it's not all so serious. Take the mathematical expression e^x and carry out the technique called integration, co-invented by Isaac Newton. Ignoring the usual constant that appears in such a calculation, you get back e^x. This standstill only happens with e^x or multiples of it.

That leads to one of the best-worst maths jokes ever. Why is e^x always stood alone at parties? Because when it tries to integrate nothing happens. *Timothy Revell*

The golden ratio

The most beautiful number ever?

YOU have probably heard of the Fibonacci sequence, that list of numbers where the next digit is given by adding the previous two. It goes 1, 1, 2, 3, 5, 8, 13 and so on. But here's something strange: work out the ratio of each number and its predecessor, and you start edging towards a specific number. Its first few digits are 1.618.

This mysterious beast is the golden ratio, and it crops up a lot. Try drawing a diagonal line connecting two vertices of a regular pentagon. Divide the length of that line by the length of the pentagon's sides and there it is. Something similar is possible with an equilateral triangle.

It turns out to be a quirk of maths. Imagine you have a number, A, and a larger one, B. If you set the numbers so that the ratio of B to A is the same as $A+B$ to B, then that ratio is always the golden ratio.

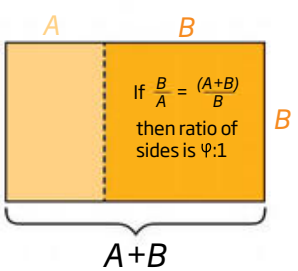
That might have been the end of the matter, but the ratio has taken on a life of its own. Search for it online and you will be inundated with claims that ancient Greek architecture or the human face exhibit such proportions, and that people find it immensely aesthetically pleasing.

The truth is murkier. The human body has countless different proportions, and some of them seem to be close to the golden ratio, but not for everyone. The ancient Greek architects were aware of the golden ratio, so it is possible that they made use of it. To find out, just measure the ruins, you might say. But then there's the question of which bits you measure – look hard enough and you will find the ratio if you want to.

A similar problem plagues studies that ask people to rate the aesthetics of artworks that incorporate the golden ratio and others that don't. It's not clear whether that judgement is really based on the ratio, or whether the association is learned or innate. Luckily, maths contains beauty enough without magic ratios. *Timothy Revell*

Golden rectangle

The rectangle below was drawn to have sides in the ratio 1.618:1. Some claim that buildings containing this "golden ratio" are especially pleasing to look at



Graham's number

The biggest number with a name of its own

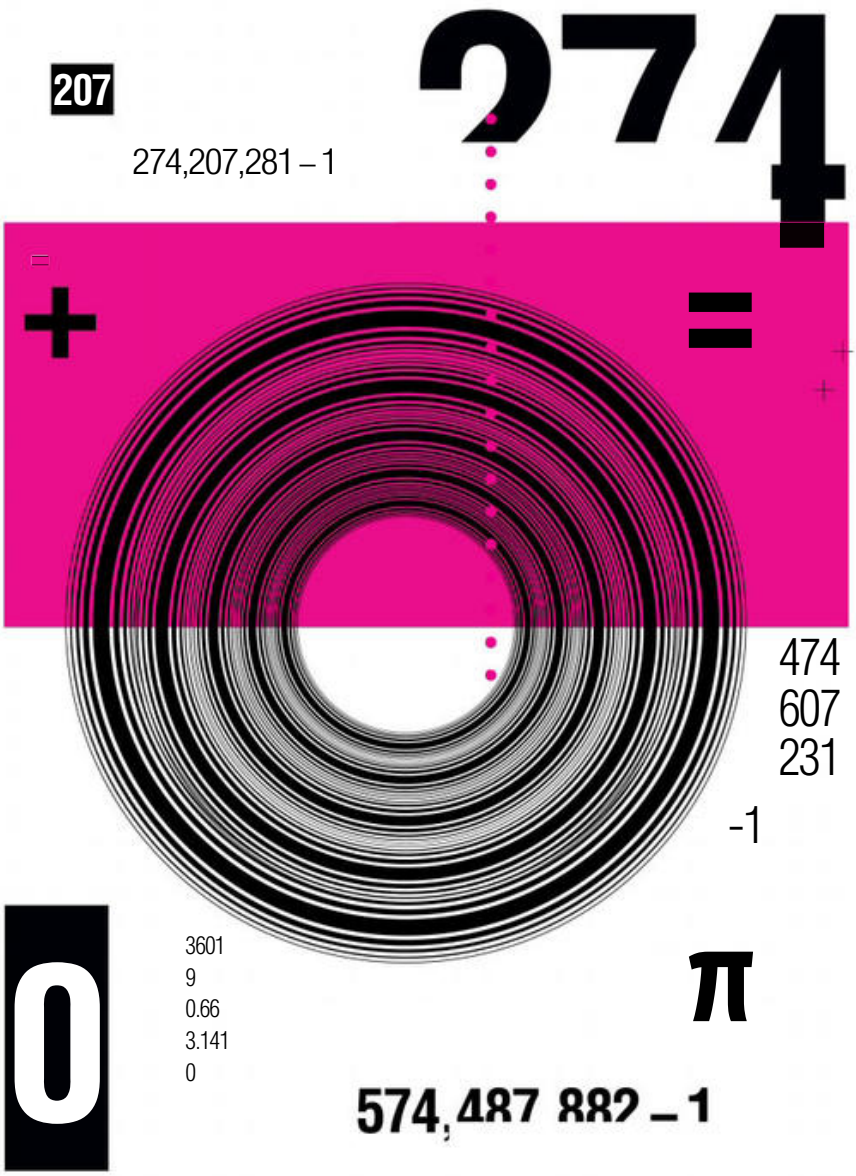
MOST numbers have never touched a human mind. There are an infinite number of numbers, after all, so it stands to reason that we have only bothered with the small ones.

But in the 1970s, Ronald Graham, a mathematician now at the University of Califonia, San Diego, was working on a puzzle that proved to have a truly gargantuan answer. He was trying to solve a problem to do with cubes in higher dimensions, and when he finally got there, the answer involved a number so large we can't write down its digits – there isn't enough space in the universe.

Yet there is a way to grasp at Graham's number. A more concise way of writing 3×3×3 is exponentiation: 3^3 means "multiply three threes together", giving 27.

We can go further, using something called Knuth's up-arrow notation. $3\uparrow 3$ means the same as 3^3, but $3\uparrow\uparrow 3$ starts a rising tower. The two arrows tell us to repeat the exponentiation, giving us 3^{3^3}, which is around 7.6 trillion.

Add a third arrow, $3\uparrow\uparrow\uparrow 3$, and things



take a major uptick, so that you reach an unimaginable stack of exponentiation upon exponentiation. Graham's number is written as 64 layers of up-arrow notation, with each layer longer than the last. In case you're wondering, its last digit is 7.

Graham's number is a whopper, but we can think of bigger ones still.

Take the function TREE(n), which relates to putting a certain number of labels on mathematical objects similar to a family tree, as part of a proof known as Kruskal's tree theorem. TREE(1) is 1. TREE(2) is 3. TREE(3) is so big it makes Graham's number seem practically zero.

Another function, called Busy Beaver, grows so fast that it has been mathematically proven to be impossible for any computer program to calculate all but its smallest values.

Busy Beaver was recently used to show that some problems are impossible to solve using the standard axioms of mathematics. But that's another big problem altogether. *Jacob Aron*

$2^{74,207,281}-1$

The force behind encryption

MULTIPLY 2 by itself just over 74 million times, then subtract 1. This is the largest known prime number, with more than 22 million digits. It is also a Mersenne prime, one equal to a power of 2, minus 1.

Other numbers in the Mersenne club include 3 and 31, but finding larger ones is no easy task. We have only discovered 49 of them.

We rely on large primes like these to ensure that all sorts of online transactions are encrypted, so that only the intended recipient can unscramble them. The idea is that the receiver multiplies two big primes to create a new number called the public key. Anyone with this key can encrypt messages, but to turn them from gobbledegook to something meaningful requires knowledge of the original two primes.

Multiplying primes together is easy for computers, but for a large answer working out the primes that produced it essentially means trying all the possibilities. That's practically impossible, making the whole process secure.

We don't really need to find a 50th Mersenne prime for the sake of encryption. But it'd be nice all the same. *Timothy Revell*

CREDIT CHECK

Ever wondered how a website knows you've typed your credit card number in wrongly before it's sent it to your bank for verification? It's down to the Luhn algorithm, brainchild of the German-born IBM engineer Hans Peter Luhn in 1954. A pioneer of mechanical data storage, Luhn was also a prolific inventor who held more than 80 patents, including one for ornamented, knitted stockings.

But the algorithm is his enduring achievement. The digits of most major credit card numbers are chosen so that, if you apply the Luhn algorithm to them, the result will be divisible by 10. Get any single digit wrong, and the number that comes out the other end won't end in a zero – and in the blink of an eye the computer says no. *Richard Webb*

CHECK YOUR CREDIT CARD NUMBER

1. Write down the 16-digit number backwards.
2. Add together all the odd digits – the ones in first, third, fifth position and so on.
3. Next, take all the digits in even positions and double them. If any of these are two-digit numbers, add the actual digits of those numbers together to get a one-digit number. Now add those numbers up.
4. Add together your answers from steps 2 and 3. The last digit of this number must be 0.

The Laplace limit

Why we don't stray far from the sun

IN 1609 the great astronomer Johannes Kepler published a book called *Astronomia Nova*. This "new astronomy" delivered a bombshell: planets revolve around the sun in ellipses, not perfect circles. But the equation at the heart of the revelations, Kepler's equation, had astronomer's heads spinning faster than the planets they were studying.

The formula describes the relationship between the coordinates of an object in orbit and the time elapsed from an arbitrary starting point. Actually solving it to find that location is fiendishly tricky.

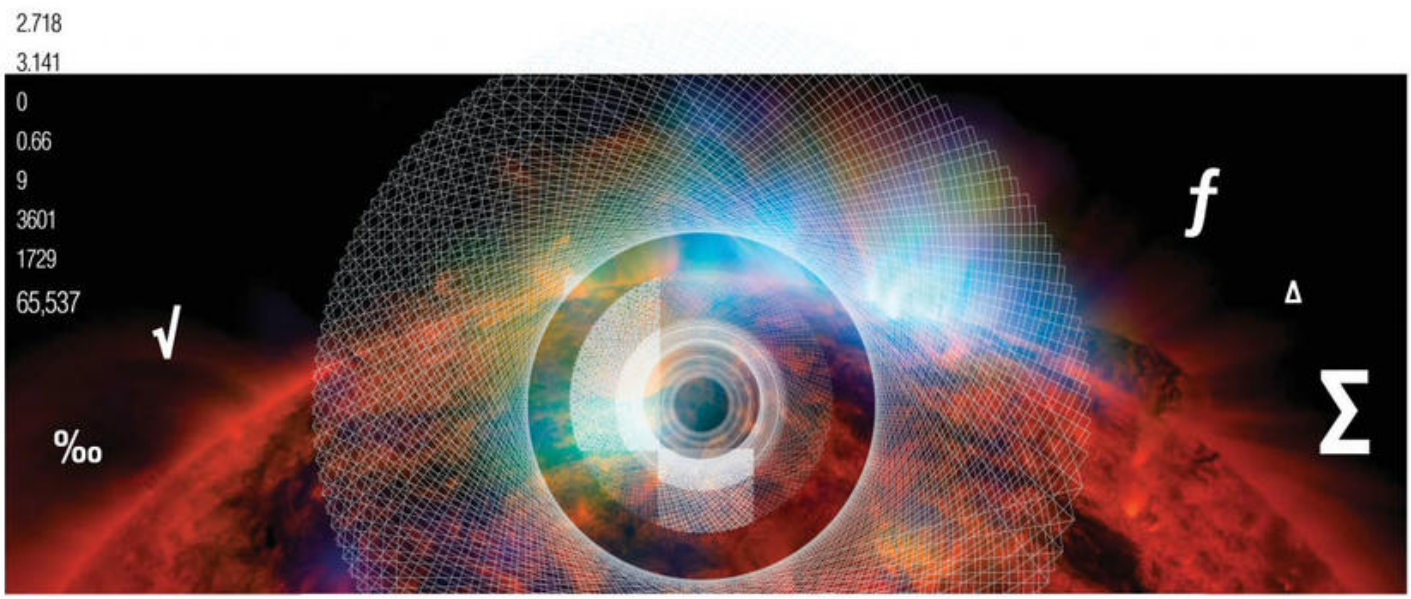
It took 150 years to find a mathematical way to solve the formula. The laborious process involved long strings of mathematical expressions known as series expansions. But the French polymath Pierre-Simon Laplace showed that this method would not work if the orbit was too elliptical.

You can quantify how far removed an ellipse is from a circle with a measure called eccentricity. A circle has an eccentricity of 0, and for values greater than that things become more skewed.

What Laplace found is that for orbits with an eccentricity of more than about 0.66 – now known as the Laplace limit – the method would not converge on a solution.

This means that "in general, orbits are less stable if the eccentricity is higher," says Gongjie Li of Harvard University. Fortunately, Earth's orbital eccentricity is about 0.02. Bodies farther out often have higher eccentricities. Pluto's is 0.25.

This doesn't mean that orbits with an eccentricity of more than 0.66 are impossible. Halley's comet has an eccentricity of 0.9. But that's best thought of as a fly-by rather than an orbit, really. The comet's swinging loop brings it close to the sun, then catapults it into the coldest reaches of the solar system. Not a place we'd want to be. *Stuart Clark*



The Lyapunov exponent

The boundary of chaos

AND the weather on Tuesday will be exponential errors, followed by a loss in predictability. Because of Lyapunov exponents, it is impossible to accurately forecast the weather more than a few days ahead. Instead of predictability, there is chaos.

In the late 19th century, the Russian mathematician Aleksandr Lyapunov invented these numbers to describe how sensitive a system is to its starting point. Imagine, for example, throwing a ball across a field. Provided you know the angle and speed at launch, you can calculate where the ball will land to a good degree of accuracy without worrying about small effects like air resitance. If your measurements of the angle are a bit off, that doesn't matter either. This situation would have a Lyapunov exponent of 0, or perhaps a negative value.

Above that threshold of zero lies unpredictability. The weather is a case in point because tiny differences in starting conditions, such as in air pressure or temperature, grow exponentially over time to cause wildly different outcomes. If throwing a ball were like this, a launch angle of 30 degrees might arrive at catching height for your friend, while an angle of 30.00000001 degrees might land the ball on the moon. Mathematicians call this chaos.

Positive Lyapunov exponents make long-term weather forecasts impossible. As we can never measure wind speed, say, with total accuracy, an initial, barely noticeable error will grow so that in only a few days the forecast will be mostly error. In countries like the UK, where air currents are highly changeable, the Lyapunov exponent of the weather is much higher than in the tropics.

"We cannot predict the future. Any little uncertainty gets amplified exponentially by chaos," says Francesco Ginelli at the University of Aberdeen, UK. Whether it is predicting the weather, the stock markets or the next president, Lyapunov exponents tell us our efforts are futile. But experience tells us we're unlikely to stop trying. *Timothy Revell*

LUCA PIERRO PHOTOGRAPHY/FLICKR SELECT/GETTY

From zero to hero

A concept of zero is essential for arithmetic to work smoothly – why then did the idea take so long to catch on? **Richard Webb** follows its convoluted path

I USED to have seven goats. I bartered three for corn; I gave one to each of my three daughters as dowry; one was stolen. How many goats do I have now?

This is not a trick question. Oddly, though, for much of human history we have not had the mathematical wherewithal to supply an answer. There is evidence of counting that stretches back five millennia in Egypt, Mesopotamia and Persia. Yet even by the most generous definition, a mathematical conception of nothing – a zero – has existed for less than half that time. Even then, the civilisations that discovered it missed its point entirely. In Europe, indifference, myopia and fear stunted its development for centuries. What is it about zero that stopped it becoming a hero?

This is a tangled story of two zeroes: zero as a symbol to represent nothing, and zero as a number that can be used in calculations and has its own mathematical properties. It is natural to think the two are the same. History teaches us something different.

Zero the symbol was in fact the first of the two to pop up by a long chalk. This is the sort of character familiar from a number such as 2012. Here it acts as a placeholder in our "positional" numerical notation, whose crucial feature is that a digit's value depends on where it is in a number. In the number 2012 a "2" crops up twice, once to mean 2 and once to mean 2000. That's because our positional system uses "base" 10 – so a move of one place to the left in a number means a digit's worth increases by a further power of 10.

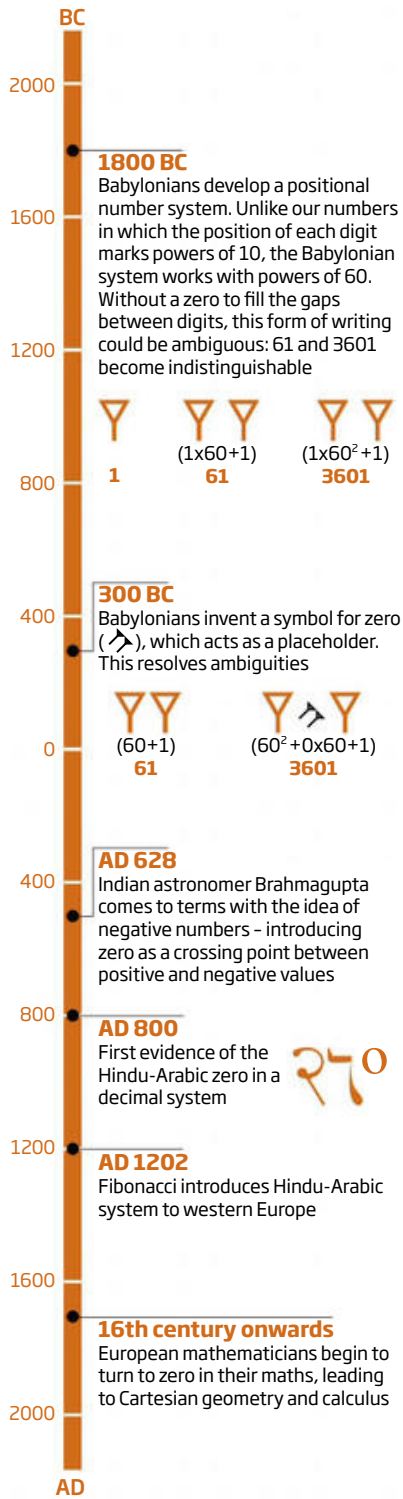
It is through such machinations that the string of digits "2012" comes to have the properties of a number with the value equal to $2 \times 10^3 + 0 \times 10^2 + 1 \times 10^1 + 2$. Zero's role is pivotal: were it not for its unambiguous presence, we might easily mistake 2012 for 212, or perhaps 20012, and our calculations could be out by hundreds or thousands.

The first positional number system was used to calculate the passage of the seasons and the years in Babylonia, modern-day Iraq, from around 1800 BC onwards. Its base was not 10, but 60. It didn't have a symbol for every whole number up to the base, unlike the "dynamic" system of digits running from 1 to 9 that is the bread and butter of our base-10 system. Instead it had just two symbols, for 1 and 10, which were clumped together in groups with a maximum headcount of 59. For example, 2012 equates to $33 \times 60^1 + 32$, and so it would have been represented by two adjacent groups of symbols: one clump of three 10s and three ones; and a second clump of three 10s and two ones.

This particular number has nothing missing. Quite generally, though, for the first 15 centuries or so of the Babylonian positional numbering system the absence of any power of 60 in the transcription of any number was marked not by a symbol, but (if you were lucky) just by a gap. What changed around 300 BC we don't know; perhaps one egregious confusion of positions too many. But it seems to have ›

A brief history of nothing

Zero is crucial for mathematics, but it has taken thousands of years for its importance to be recognised



been at around this time that a third symbol, a curious confection of two left-slanting arrows (see timeline, overleaf), started to fill missing places in the stargazers' calculations.

This was the world's first zero. Some seven centuries later, on the other side of the world, it was invented a second time. Mayan priest-astronomers in central America began to use a snail-shell-like symbol to fill gaps in the (almost) base-20 positional "long-count" system they used to calculate their calendar.

Zero as a placeholder was clearly a useful concept, then. It is a frustration entirely typical of zero's vexed history, though, that neither the Babylonians nor the Mayans realised quite how useful it could be.

In any dynamic, positional number system, a placeholder zero assumes almost unannounced a new guise: it becomes a mathematical "operator" that brings the full power of the system's base to bear. This becomes obvious when we consider the result of adding a placeholder zero to the end of a decimal number string. The number 2012 becomes 20120, magically multiplied by the base of 10. We intuitively take advantage of this characteristic whenever we sum two or more numbers, and the total of a column ticks over from 9 to 10. We "carry the one" and leave a zero to ensure the right answer. The simplicity of such algorithms is the source of our system's supple muscularity in manipulating numbers.

Facing the void

We shouldn't blame the Babylonians or Mayans for missing out on such subtlety: various blemishes in their numerical systems made it hard to spot. And so, although they found zero the symbol, they missed zero the number.

Zero is admittedly not an entirely welcome addition to the pantheon of numbers. Accepting it invites all sorts of logical wrinkles that, if not handled with due care and attention, can bring the entire number system crashing down. Adding zero to itself does not result in any increase in its size, as it does for any other number. Multiply any number, however big, by zero and it collapses down to zero. And let's not even delve into what happens when we divide a number by zero.

Classical Greece, the next civilisation to handle the concept, was certainly not keen to tackle zero's complexities. Greek thought was wedded to the idea that numbers expressed geometrical shapes; and what shape would correspond to something that wasn't there? It could only be the total absence of something,

"Sitting at the threshold between the positive and negative worlds was sunya, the nothingness"

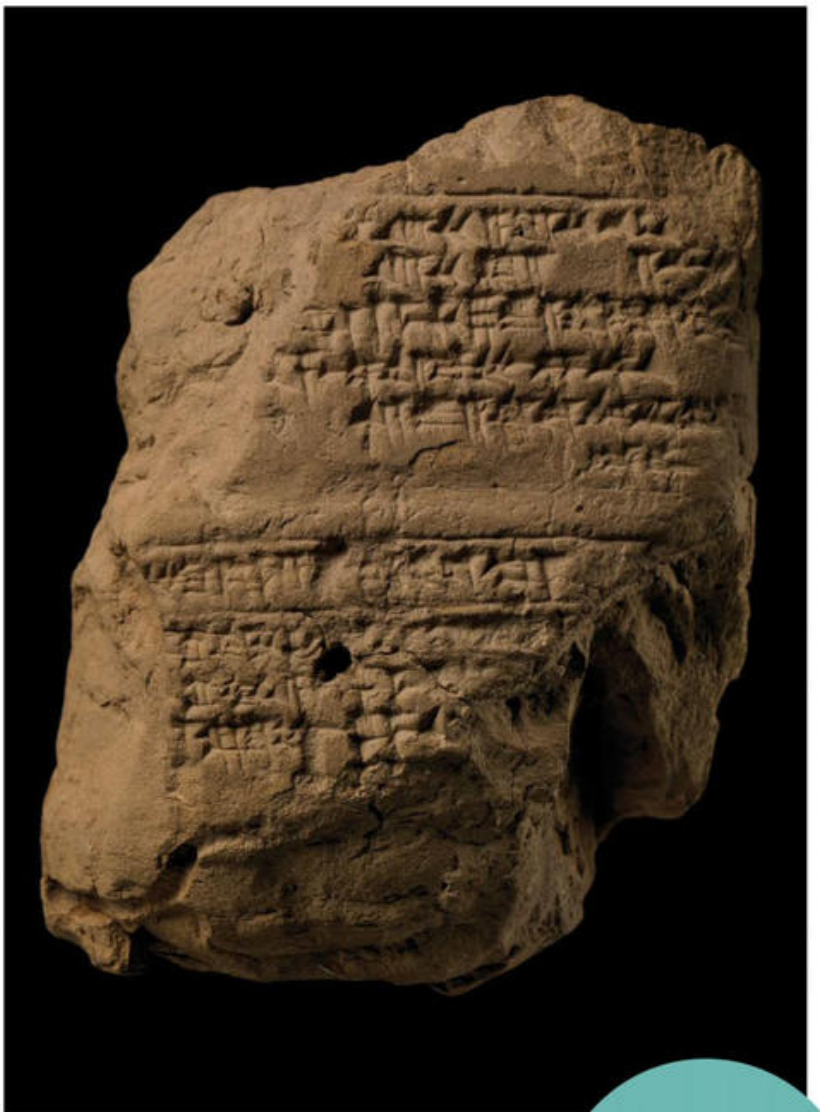
the void – a concept that the dominant cosmology of the time had banished.

Largely the product of Aristotle and his disciples, this world view saw the planets and stars as embedded in a series of concentric celestial spheres of finite extent. These spheres were filled with an ethereal substance, all centred on Earth and set in motion by an "unmoved mover". It was a picture later eagerly co-opted by Christian philosophy, which saw in the unmoved mover a ready-made identity for God. And since there was no place for a void in this cosmology, it followed that it – and everything associated with it – was a godless concept.

Eastern philosophy, rooted in ideas of eternal cycles of creation and destruction, had no such qualms. And so the next great staging post in zero's journey was not to Babylon's west, but to its east. It is found in *Brahmasphutasiddhanta*, a treatise on the relationship of mathematics to the physical world written in India in around AD 628 by the astronomer Brahmagupta.

Brahmagupta was the first person we see treating numbers as purely abstract quantities separate from any physical or geometrical reality. This allowed him to consider unorthodox questions that the Babylonians and Greeks had ignored or dismissed, such as what happens when you subtract from one number a number of greater size. In geometrical terms this is a nonsense: what area is left when a larger area is subtracted? Equally, how could I ever have sold or bartered more goats than I had in the first place? As soon as numbers become abstract entities, however, a whole new world of possibilities is opened up – the world of negative numbers.

The result was a continuous number line stretching as far as you could see in both directions, showing both positive and negative numbers. Sitting in the middle of this line, a distinct point along it at the threshold between the positive and negative worlds, was *sunya*, the nothingness. Indian mathematicians had dared to look into the



Early Babylonian calculations on their version of a tablet PC

THE BRITISH MUSEUM/BRITISH MUSEUM IMAGES

void – and a new number had emerged.

It was not long before they unified this new number with zero the symbol. Recent dating of a manuscript held at the Bodleian Library in Oxford, UK, suggests that as early as the 3rd or 4th century AD Hindu mathematicians were using a squashed-egg symbol recognisably close to our own zero as a placeholder. Brahmagupta's innovation made this placeholder zero a full member of a dynamic positional number system running from 0 to 9. It marked the birth of the purely abstract number system now used throughout the world, and soon spawned a new way of doing mathematics to go with it: algebra.

News of these innovations took a long time to filter through to Europe. It was only in 1202 that a young Italian, Leonardo of Pisa – better remembered as Fibonacci – published a book, *Liber Abaci*, in which he presented details of the Arabic counting system he had encountered on a journey to the Mediterranean's southern shores, and demonstrated the superiority of this notation over the abacus for the deft performance of complex calculations.

"The Babylonians found zero the symbol, but missed zero the number"

While merchants and bankers were quickly convinced of the Hindu-Arabic system's usefulness, the governing authorities were less enamoured. In 1299, the city of Florence, Italy, banned the use of the Hindu-Arabic numerals, including zero. They considered the ability to inflate a number's value hugely simply by adding a digit on the end – a facility not available in the then-dominant, non-positional system of Roman numerals – to be an open invitation to fraud.

Zero the number had an even harder time. Schisms, upheavals, reformation and counter-reformation in the church meant a continuing debate as to the worth of Aristotle's ideas about the cosmos, and with it the orthodoxy or otherwise of the void. Only the Copernican revolution – the crystal-sphere-shattering revelation that Earth moves around the sun – began, slowly, to shake European mathematics free of the shackles of Aristotelian cosmology from the 16th century onwards.

By the 17th century, the scene was set for zero's final triumph. It is hard to point to a single event that marked it. Perhaps it was the advent of the coordinate system invented by the French philosopher and mathematician René Descartes. His Cartesian system married algebra and geometry to give every geometrical shape a new symbolic representation with zero, the unmoving heart of the coordinate system, at its centre. Zero was far from irrelevant to geometry, as the Greeks had suggested: it was essential to it. Soon afterwards, the new tool of calculus showed that you had first to appreciate how zero merged into the infinitesimally small to explain how anything in the cosmos could change its position at all – a star, a planet, a hare overtaking a tortoise. Zero was itself the prime mover.

Thus a better understanding of zero became the fuse of the scientific revolution that followed. Subsequent events have confirmed just how essential zero is to mathematics and all that builds on it (see "Nothing in common", page 36). Looking at zero sitting quietly in a number today, and primed with the concept from a young age, it is equally hard to see how it could ever have caused so much confusion and distress. A case, most definitely, of much ado about nothing. ■

{

Nothing in common

A collection of nothings means everything to mathematics, as **Ian Stewart** explains

THE mathematicians' version of nothing is the empty set. This is a collection that doesn't actually contain anything, such as my own collection of vintage Rolls-Royces. The empty set may seem a bit feeble, but appearances deceive; it provides a vital building block for the whole of mathematics.

It all started in the late 1800s. While most mathematicians were busy adding a nice piece of furniture, a new room, even an entire storey to the growing mathematical edifice, a group of worrywarts started to fret about the cellar. Innovations like non-Euclidean geometry and Fourier analysis were all very well – but were the underpinnings sound? To prove they were, a basic idea needed sorting out that no one really understood. Numbers.

Sure, everyone knew how to do sums. Using numbers wasn't the problem. The big question was what they were. You can show someone two sheep, two coins, two albatrosses, two galaxies. But can you show them two?

The symbol "2"? That's a notation, not the number itself. Many cultures use a different symbol. The word "two"? No, for the same reason: in other languages it might be *deux* or *zwei* or *futatsu*. For thousands of years humans had been using numbers to great effect; suddenly a few deep thinkers realised no one had a clue what they were.

An answer emerged from two different lines of thought: mathematical logic, and Fourier analysis, in which a complex waveform describing a function is represented as a combination of simple sine waves. These two areas converged on one idea. Sets.

A set is a collection of mathematical objects – numbers, shapes, functions, networks, whatever. It is defined by listing or characterising its members. "The set with members 2, 4, 6, 8" and "the set of even integers between 1 and 9" both define the same set, which can be written as {2, 4, 6, 8}.

Around 1880 the mathematician Georg Cantor developed an extensive theory of sets. He had been trying to sort out some technical issues in Fourier analysis related to discontinuities – places where the waveform makes sudden jumps. His answer involved the structure of the set of discontinuities. It wasn't the individual discontinuities that mattered, it was the whole class of discontinuities.

How many dwarfs?

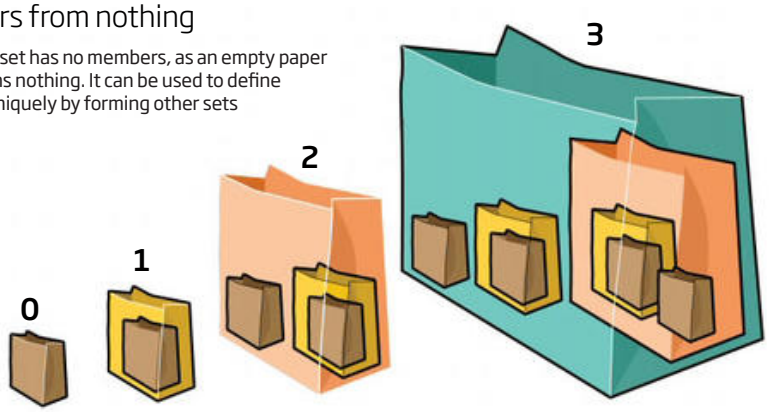
One thing led to another. Cantor devised a way to count how many members a set has, by matching it in a one-to-one fashion with a standard set. Suppose, for example, the set is {Doc, Grumpy, Happy, Sleepy, Bashful, Sneezy, Dopey}. To count them we chant "1, 2, 3…" while working along the list: Doc (1), Grumpy (2), Happy (3), Sleepy (4), Bashful (5), Sneezy (6) Dopey (7). Right: seven dwarfs. We can do the same with the days of the week: Monday (1), Tuesday (2), Wednesday (3), Thursday (4), Friday (5), Saturday (6), Sunday (7).

Another mathematician of the time, Gottlob Frege, picked up on Cantor's ideas and thought they could solve the big philosophical problem of numbers. The way to define them, he believed, was through the deceptively simple process of counting.

What do we count? A collection of things – a set. How do we count it? By matching the things in the set with a standard set of known size. The next step was simple but devastating: throw away the numbers. You could use the dwarfs to count the days of the week. Just set up the correspondence: Monday (Doc), Tuesday (Grumpy)… Sunday (Dopey). There are Dopey days in the week. It's a perfectly reasonable alternative number system. It doesn't (yet) tell us what a number is, but it gives a way to define "same number". The number of days equals the number of dwarfs,

Numbers from nothing

The empty set has no members, as an empty paper bag contains nothing. It can be used to define numbers uniquely by forming other sets



not because both are seven, but because you can match days to dwarfs.

What, then, is a number? Mathematical logicians realised that to define the number 2, you need to construct a standard set which intuitively has two members. To define 3, use a standard set with three numbers, and so on. But which standard sets to use? They have to be unique, and their structure should correspond to the process of counting. This was where the empty set came in and solved the whole thing by itself.

Zero is a number, the basis of our entire number system (see "From zero to hero", page 33). So it ought to count the members of a set. Which set? Well, it has to be a set with no members. These aren't hard to think of: "the set of all honest bankers", perhaps, or "the set of all mice weighing 20 tonnes". There is also a mathematical set with no members: the empty set. It is unique, because all empty sets have exactly the same members: none. Its symbol, introduced in 1939 by a group of mathematicians that went by the pseudonym Nicolas Bourbaki, is Ø. Set theory needs Ø for the same reason that arithmetic needs 0: things are a lot simpler if you include it. In fact, we can define the number 0 as the empty set.

What about the number 1? Intuitively, we need a set with exactly one member. Something unique. Well, the empty set is unique. So we define 1 to be the set whose only member is the empty set: in symbols, {Ø}. This is not the same as the empty set, because it has one member, whereas the empty set has none. Agreed, that member happens to be the empty set, but there is one of it. Think of a set as a paper bag containing its members. The empty set is an empty paper bag. The set whose only member is the empty set is a paper bag containing an empty paper bag. Which is different: it's got a bag in it (see diagram above).

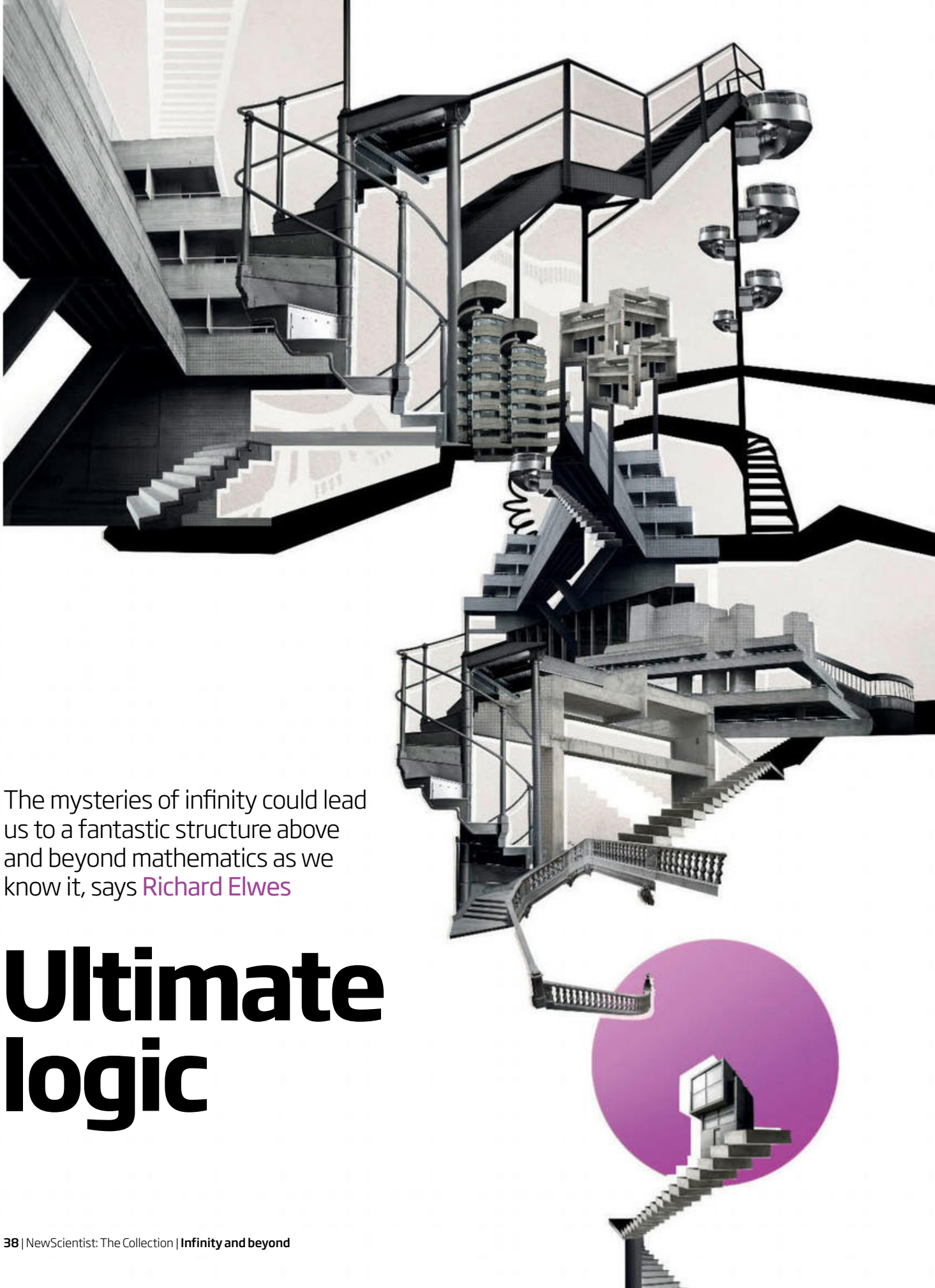
The key step is to define the number 2. We need a uniquely defined set with two members. So why not use the only two sets we've mentioned so far: Ø and {Ø}? We therefore define 2 to be the set {Ø, {Ø}}. Which, thanks to our definitions, is the same as {0, 1}.

Now a pattern emerges. Define 3 as {0, 1, 2}, a set with three members, all of them already defined. Then 4 is {0, 1, 2, 3}, 5 is {0, 1, 2, 3, 4}, and so on. Everything traces back to the empty set: for instance, 3 is {Ø, {Ø}, {Ø, {Ø}}} and 4 is {Ø, {Ø}, {Ø, {Ø}}, {Ø, {Ø}, {Ø, {Ø}}}}. You don't want to see what the number of dwarfs looks like.

The building materials here are abstractions: the empty set and the act of forming a set by listing its members. But the way these sets relate to each other leads to a well-defined construction for the number system, in which each number is a specific set that intuitively has that number of members. The story doesn't stop there. Once you have defined the positive whole numbers, similar set-theoretic trickery defines negative numbers, fractions, real numbers (infinite decimals), complex numbers... all the way to the latest fancy mathematical concept in quantum theory or whatever.

So now you know the dreadful secret of mathematics: it's all based on nothing. ■

"The empty set is the set with nothing in it. The number 1 is the set with only the empty set in it. And so on"



The mysteries of infinity could lead us to a fantastic structure above and beyond mathematics as we know it, says Richard Elwes

Ultimate logic

CHAPTER FOUR

INFINITY

WHEN David Hilbert left the podium at the Sorbonne in Paris, France, on 8 August 1900, few of the assembled delegates seemed overly impressed. According to one contemporary report, the discussion following his address to the second International Congress of Mathematicians was "rather desultory". Passions seem to have been more inflamed by a subsequent debate on whether Esperanto should be adopted as mathematics' working language.

Yet Hilbert's address set the mathematical agenda for the 20th century. It crystallised into a list of 23 crucial unanswered questions, including how to pack spheres to make best use of the available space, and whether the Riemann hypothesis, which concerns how the prime numbers are distributed, is true.

Today many of these problems have been resolved, sphere-packing among them. Others, such as the Riemann hypothesis, have seen little or no progress. But the first item on Hilbert's list stands out for the sheer oddness of the answer supplied by generations of mathematicians since: that mathematics is simply not equipped to provide an answer.

This curiously intractable riddle is known as the continuum hypothesis, and it concerns that most enigmatic quantity, infinity. In 2010, at that same forum Hilbert addressed, the International Congress of Mathematicians, this time held in Hyderabad, India, a respected US mathematician claimed to have cracked it. He arrived at the solution not by using mathematics as we know it, but by building a new, radically stronger logical structure: a structure he dubs "ultimate L".

The journey to this point began in the early 1870s, when the German Georg Cantor was laying the foundations of set theory. Set theory deals with the counting and manipulation of collections of objects, and provides the crucial logical underpinnings of mathematics: because numbers can be associated with the size of sets, the rules for manipulating sets also determine the logic of arithmetic and everything that builds on it.

These dry, slightly insipid logical considerations gained a new tang when Cantor asked a critical question: how big can sets get? The obvious answer – infinitely big – turned out to have a shocking twist: infinity is not one entity, but comes in many levels.

How so? You can get a flavour of why by counting up the set of whole numbers: 1, 2, 3, 4, 5… How far can you go? Why, infinitely far, of course – there is no biggest whole number. This is one sort of infinity, the smallest, "countable" level, where the action of arithmetic takes place.

Now consider the question "how many points are there on a line?" A line is perfectly straight and smooth, with no holes or gaps; it contains infinitely many points. But this is not the countable infinity of the whole numbers, where you bound upwards in a series of defined, well-separated steps. This is a smooth, continuous infinity that describes geometrical objects. It is characterised not by the whole numbers, but by the real numbers: the whole numbers plus all the numbers in between that have as many decimal places as you please – 0.1, 0.01, $\sqrt{2}$, π and so on.

Cantor showed that this "continuum" infinity is in fact infinitely bigger than the countable, whole-number variety. What's more, it is merely a step in a staircase leading to ever-higher levels of infinities stretching up as far as, well, infinity.

While the precise structure of these higher infinities remained nebulous, a more immediate question frustrated Cantor. Was there an intermediate level between the countable infinity and the continuum? He suspected not, but was unable to prove it. His hunch about the non-existence of this mathematical mezzanine became known as the continuum hypothesis.

Attempts to prove or disprove the continuum hypothesis depend on analysing all possible infinite subsets of the real numbers. If every one is either countable or has the same size as the full continuum, then it is correct. Conversely, even one subset of intermediate size would render it false.

A similar technique using subsets of the ›

EMMA-MARIE EMMERMANN

whole numbers shows that there is no level of infinity below the countable. Tempting as it might be to think that there are half as many even numbers as there are whole numbers in total, the two collections can in fact be paired off exactly. Indeed, every set of whole numbers is either finite or countably infinite.

Applied to the real numbers, though, this approach bore little fruit, for reasons that soon became clear. In 1885, the Swedish mathematician Gösta Mittag-Leffler had blocked publication of one of Cantor's papers on the basis that it was "about 100 years too soon". And as the British mathematician and philosopher Bertrand Russell showed in 1901, Cantor had indeed jumped the gun. Although his conclusions about infinity were sound, the logical basis of his set theory was flawed, resting on an informal and ultimately paradoxical conception of what sets are.

It was not until 1922 that two German mathematicians, Ernst Zermelo and Abraham Fraenkel, devised a series of rules for manipulating sets that was seemingly robust enough to support Cantor's tower of infinities and stabilise the foundations of mathematics. Unfortunately, though, these rules delivered no clear answer to the continuum hypothesis. In fact, they seemed strongly to suggest there might even not be an answer.

"A Swedish mathematician once blocked publication of one of Cantor's papers on the basis that it was 'about 100 years too soon'"

Agony of choice

The immediate stumbling block was a rule known as the "axiom of choice". It was not part of Zermelo and Fraenkel's original rules, but was soon bolted on when it became clear that some essential mathematics, such as the ability to compare different sizes of infinity, would be impossible without it.

The axiom of choice states that if you have a collection of sets, you can always form a new set by choosing one object from each of them. That sounds anodyne, but it comes with a sting: you can dream up some twisted initial sets that produce even stranger sets when you choose one element from each. The Polish mathematicians Stefan Banach and Alfred Tarski soon showed how the axiom could be used to divide the set of points defining a spherical ball into six subsets which could then be slid around to produce two balls of the same size as the original. That was a symptom of a fundamental problem: the axiom allowed peculiarly perverse sets of real numbers to exist whose properties could never be determined. If so, this was a grim portent for ever proving the continuum hypothesis.

This news came at a time when the concept of "unprovability" was just coming into vogue. In 1931, the Austrian logician Kurt Gödel proved his notorious "incompleteness theorem". It shows that even with the most tightly knit basic rules, there will always be statements about sets or numbers that mathematics can neither verify nor disprove.

At the same time, though, Gödel had a crazy-sounding hunch about how you might fill in most of these cracks in mathematics' underlying logical structure: you simply build more levels of infinity on top of it. That goes against anything we might think of as a sound building code, yet Gödel's guess turned out to be inspired. He proved his point in 1938. By starting from a simple conception of sets compatible with Zermelo and Fraenkel's rules and then carefully tailoring its infinite superstructure, he created a mathematical environment in which both the axiom of choice and the continuum hypothesis are simultaneously true. He dubbed his new world the "constructible universe" – or simply "L".

L was an attractive environment in which to do mathematics, but there were soon reasons to doubt it was the "right" one. For a start, its infinite staircase did not extend high enough to fill in all the gaps known to exist in the underlying structure. In 1963 Paul Cohen of Stanford University in California put things into context when he developed a method for producing a multitude of mathematical universes to order, all of them compatible with Zermelo and Fraenkel's rules.

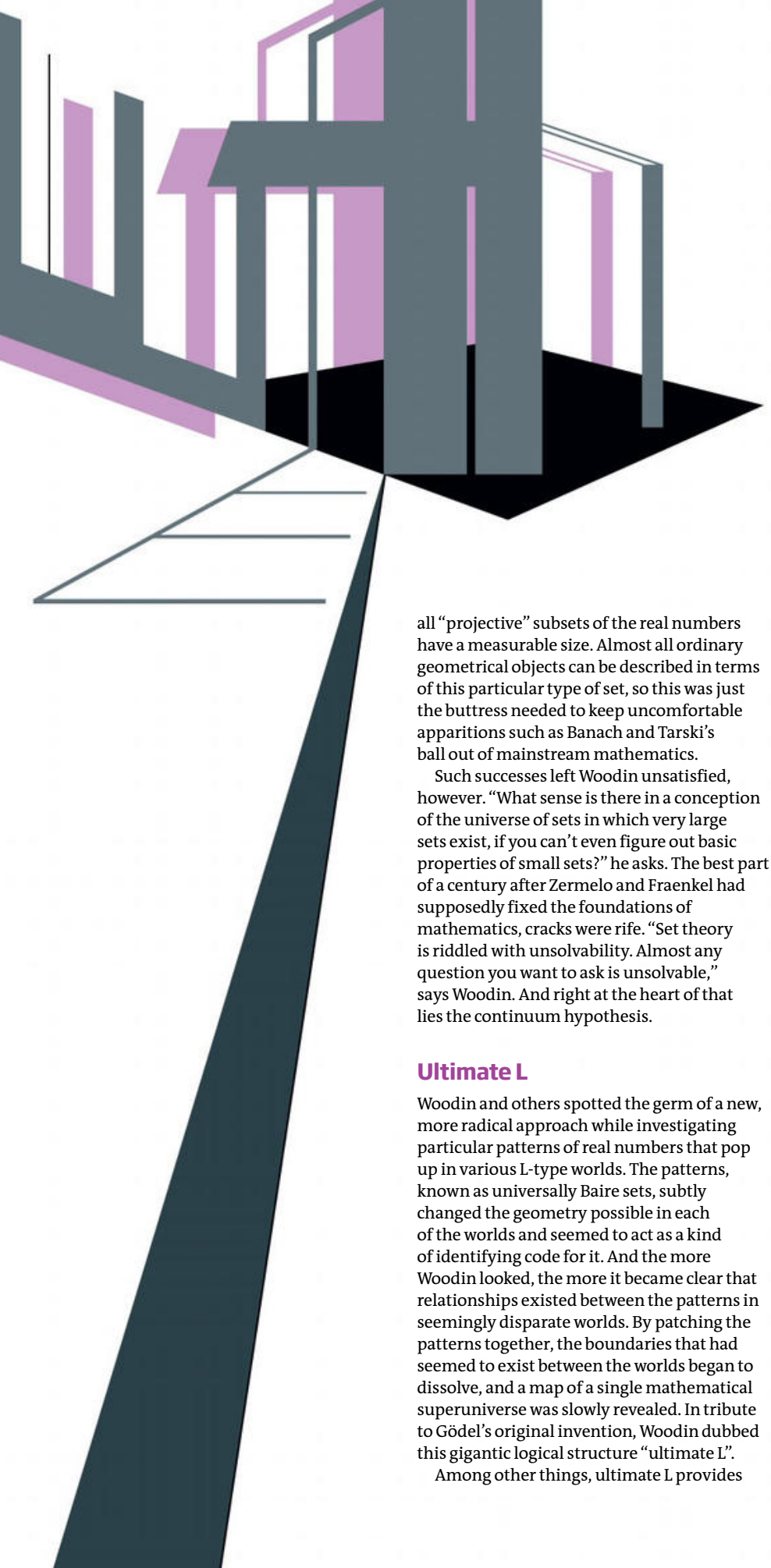
This was the beginning of a construction boom. "Over the past half-century, set theorists have discovered a vast diversity of models of set theory, a chaotic jumble of set-theoretic possibilities," says Joel Hamkins at the City University of New York. Some are "L-type worlds" with superstructures like Gödel's L, differing only in the range of extra levels of infinity they contain; others have wildly varying architectural styles with completely different levels and infinite staircases leading in all sorts of directions.

For most purposes, life within these structures is the same: most everyday mathematics does not differ between them, and nor do the laws of physics. But the existence of this mathematical "multiverse" also seemed to dash any notion of ever getting to grips with the continuum hypothesis. As Cohen was able to show, in some logically possible worlds the hypothesis is true and there is no intermediate level of infinity between the countable and the continuum; in others, there is one; in still others, there are infinitely many. With mathematical logic as we know it, there is simply no way of finding out which sort of world we occupy.

That's where Hugh Woodin of Harvard University first made his suggestion back in 2010. The answer, he says, can be found by stepping outside our conventional mathematical world and moving on to a higher plane.

Woodin is no "turn on, tune in" guru. A highly respected set theorist, he has already achieved his subject's ultimate accolade: a level on the infinite staircase named after him. This level, which lies far higher than anything envisaged in Gödel's L, is inhabited by gigantic entities known as Woodin cardinals.

Woodin cardinals illustrate how adding penthouse suites to the structure of mathematics can solve problems on less rarefied levels below. In 1988 the American mathematicians Donald Martin and John Steel showed that if Woodin cardinals exist, then



all "projective" subsets of the real numbers have a measurable size. Almost all ordinary geometrical objects can be described in terms of this particular type of set, so this was just the buttress needed to keep uncomfortable apparitions such as Banach and Tarski's ball out of mainstream mathematics.

Such successes left Woodin unsatisfied, however. "What sense is there in a conception of the universe of sets in which very large sets exist, if you can't even figure out basic properties of small sets?" he asks. The best part of a century after Zermelo and Fraenkel had supposedly fixed the foundations of mathematics, cracks were rife. "Set theory is riddled with unsolvability. Almost any question you want to ask is unsolvable," says Woodin. And right at the heart of that lies the continuum hypothesis.

Ultimate L

Woodin and others spotted the germ of a new, more radical approach while investigating particular patterns of real numbers that pop up in various L-type worlds. The patterns, known as universally Baire sets, subtly changed the geometry possible in each of the worlds and seemed to act as a kind of identifying code for it. And the more Woodin looked, the more it became clear that relationships existed between the patterns in seemingly disparate worlds. By patching the patterns together, the boundaries that had seemed to exist between the worlds began to dissolve, and a map of a single mathematical superuniverse was slowly revealed. In tribute to Gödel's original invention, Woodin dubbed this gigantic logical structure "ultimate L".

Among other things, ultimate L provides for the first time a definitive account of the spectrum of subsets of the real numbers: for every forking point between worlds that Cohen's methods open up, only one possible route is compatible with Woodin's map. In particular it implies Cantor's hypothesis to be true, ruling out anything between countable infinity and the continuum. If true, that would not only solve a problem bugging mathematicians for almost a century and a half, but also mark a personal turnaround for Woodin: in earlier years, he argued that the continuum hypothesis should be considered false.

Ultimate L does not rest there. Its wide, airy space allows extra steps to be bolted to the top of the infinite staircase as necessary to fill in gaps below, making good on Gödel's hunch about rooting out the unsolvability that riddles mathematics. Gödel's incompleteness theorem would not be dead, but you could chase it as far as you pleased up the staircase into the infinite attic of mathematics.

The prospect of finally removing the logical incompleteness that has bedevilled even basic areas such as number theory is enough to get many mathematicians salivating. But the jury is split on whether ultimate L is the ultimate answer.

Andrés Caicedo, a logician at Boise State University in Idaho, is cautiously optimistic. "It would be reasonable to say that this is the 'correct' way of going about completing the rules of set theory," he says. "But there are still several technical issues to be clarified before saying confidently that it will succeed."

Others are less convinced. Hamkins, who is a former student of Woodin's, holds to the idea that there simply are as many legitimate logical constructions for mathematics as we have found so far. He thinks mathematicians should learn to embrace the diversity of the mathematical multiverse, with spaces where the continuum hypothesis is true and others where it is false. The choice of which space to work in would then be a matter of personal taste and convenience. "The answer consists of our detailed understanding of how the continuum hypothesis both holds and fails throughout the multiverse," he says.

Woodin's ideas need not put paid to this choice entirely, though: aspects of many of these diverse universes will survive inside ultimate L. "One goal is to show that any universe attainable by means we can currently foresee can be obtained from the theory," says Caicedo. "If so, then ultimate L is all we need." ■

COMPLETE YOUR COLLECTION

Missed a copy of *New Scientist: The Collection*? Past issues are available to buy through our online store: **newscientist.com/thecollection**

New Scientist

HOW TO THINK ABOUT

INFINITY

IAN STEWART has an easy, if not particularly helpful, way of envisaging infinity. "I generally think of it as: (a) very big, but (b) bigger than that," says the mathematician from the University of Warwick in the UK. "When something is infinite, there is always some spare room around to put things in."

Infinity is one of those things with a preprogrammed boggle factor. Mathematically, it started off as a way of expressing the fact that some things, like counting, have no obvious end. Count to 146 and there's 147; count to a trillion and say hello to a trillion and one.

There are two ways of dealing with this, says Stewart. "You can sum it up boldly as 'there are infinitely many numbers'. But if you want to be more cautious, you just say 'there is no largest number'."

Only in the late 19th century did mathematicians plump for the first option, and begin to handle infinity as an object with properties all of its own. The key was set theory, a new way of thinking of numbers as bundles of things. The set of all whole numbers, for example, is a well-defined and unique object, and it has a size: infinity.

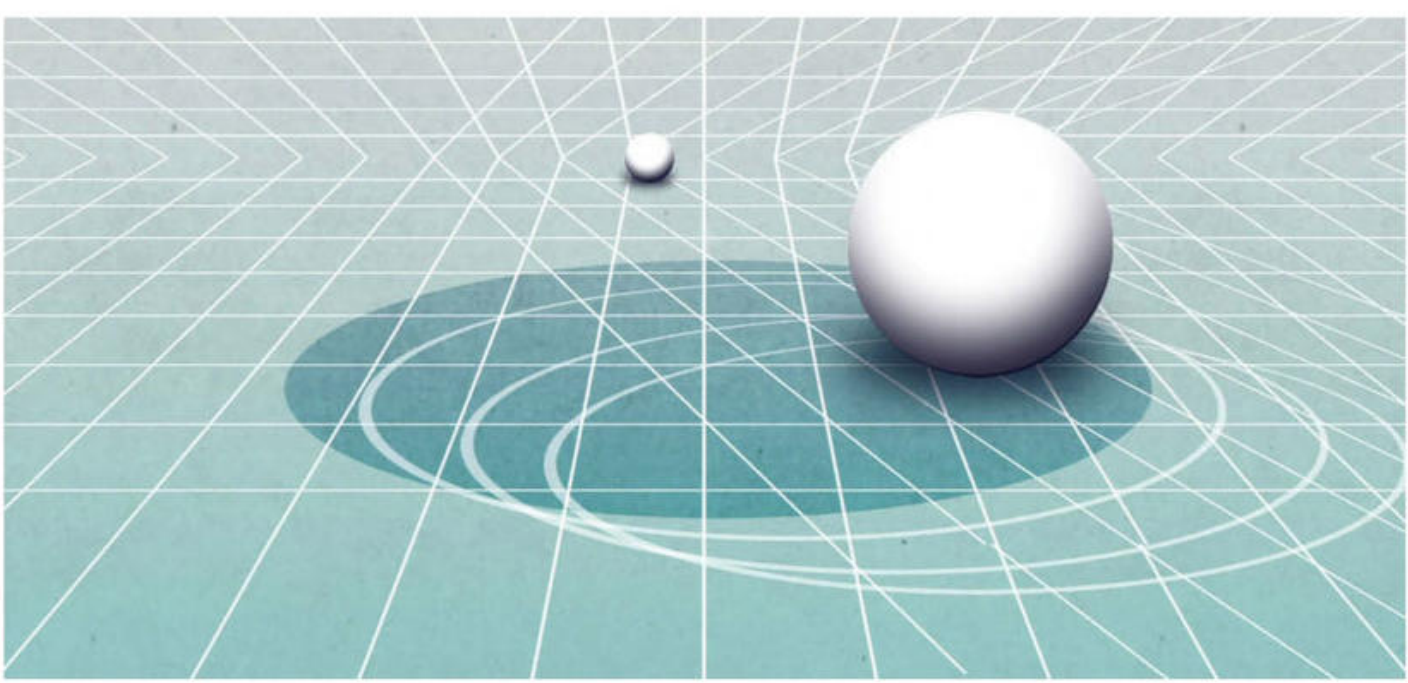
The sting in the tail, as Georg Cantor showed, is that by this definition there is more than one infinity. The set of the whole numbers defines one low-lying sort, known as countable infinity. But add in all the numbers in between, with as many decimal places as you please, and you get a smoother, more continuous infinity – one defined by a set that is infinitely bigger.

That is just the beginning. The "Woodin cardinals" proposed by Harvard University set theorist represent even more vertiginous levels of infinity. "They are so large you can't deduce their existence," says Woodin.

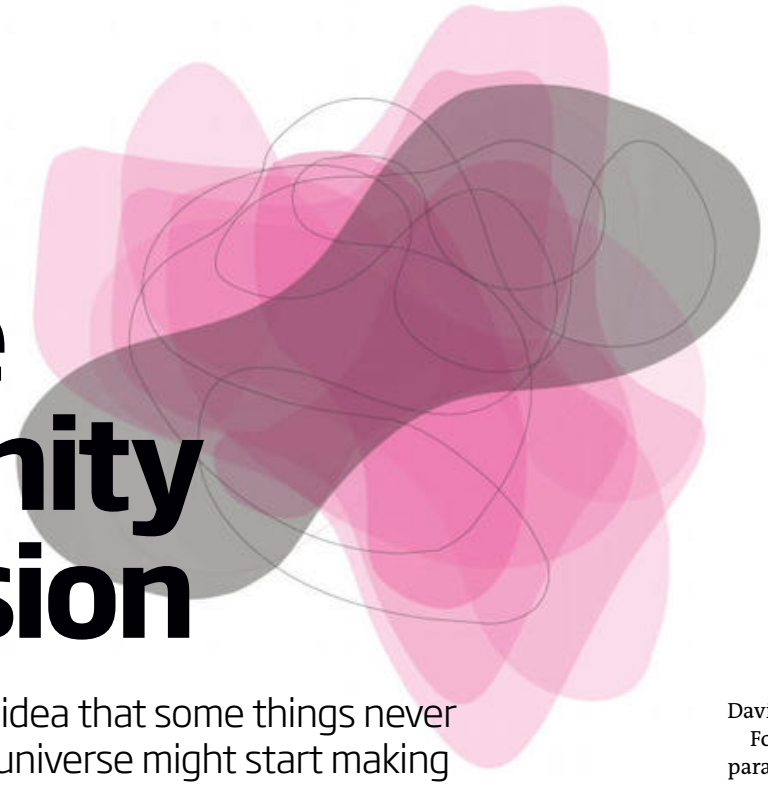
Such infinities may help solve otherwise unsolvable problems in less rarefied mathematical landscapes below (see "Ultimate logic", page 38). But they are the ultimate abstraction: although you can manipulate them logically, you can't write formulae incorporating them or devise computer programs to test predictions about them. Woodin's notepads consist mainly of cryptic marks he uses to focus his attention, to the occasional consternation of fellow plane passengers. "If they don't try to change seats, they ask me if I'm an artist," he says.

How closely our common-sense conception of endlessness matches the mathematical infinities isn't clear. But if we can't quite grasp boundarylessness, it probably doesn't matter, says Woodin – however you slice it, infinity seems far removed from anything we see in the real world.

So perhaps those enigmatic markings aren't so different from those of his fellow passengers after all. "It might be we're just playing a game," says Woodin. "Perhaps we are just doing some glorified sudoku puzzle." **Richard Webb**



MARIO WAGNER



The infinity illusion

Abandon the idea that some things never end, and the universe might start making more sense, says **Amanda Gefter**

INFINITY is a concept that defies imagination. We have a hard-enough time trying to wrap our minds around things that are merely extremely big: our solar system, our galaxy, the observable universe. But those scales are nothing compared with the infinite. Just thinking about it can make you queasy.

But we cannot avoid it. Mathematics as we know it is riddled with infinities. The number line stretches to eternity and beyond, and is infinitely divisible: countless more numbers lurk between any two others. The number of digits in a constant like π is limitless. Whether geometry, trigonometry or calculus, the mathematical manipulations we use to make sense of the world are built on the idea that some things never end.

Trouble is, once unleashed, these infinities are wild, unruly beasts. They blow up the equations with which physicists attempt to explain nature's fundamentals. They obstruct a unified view of the forces that shape the cosmos. Worst of all, add infinities to the explosive mixture that made up the infant universe and they prevent us from making any scientific predictions at all.

All of which encourages a bold speculation among a few physicists and mathematicians: can we do away with infinity?

Belief in the never-ending has not always been a mainstream view. "For most of the history of mathematics, infinity was kept at arm's length," says mathematician Norman Wildberger of the University of New South Wales in Sydney, Australia. For greats of the subject, from Aristotle to Newton and Gauss, the only infinity was a "potential" infinity. This type of infinity allows us to add 1 to any number without fear of hitting the end of the number line, but is never actually reached itself. That is a long way from accepting "actual" infinity – one that has already been reached and conveniently packaged as a mathematical entity we can manipulate in equations.

Things changed in the late 19th century, with the invention by Georg Cantor of set theory, the underpinning of modern number theory. He argued that sets containing an infinite number of elements were themselves mathematical objects. This masterstroke allowed the meaning of numbers to be pinned down in a rigorous way that had long eluded mathematicians. Within set theory, the infinite continuum of the "real" numbers, including all the rational numbers (those, like ½, which can be expressed as a ratio of integers) and the irrational numbers (those that cannot, like π) came to be treated as actual, rather than potential, infinities. "No one shall expel us from the paradise Cantor has created," the mathematician David Hilbert later declared.

For physicists, however, the infinite paradise has become more like purgatory. To take one example, the standard model of particle physics was long beset by pathological infinities, for instance in quantum electrodynamics, the quantum theory of the electromagnetic force. It initially showed the mass and charge of an electron to be infinite.

Decades of work, rewarded by many a Nobel prize, banished these nonsensical infinities – or most of them. Gravity has notoriously resisted unification with the other forces of nature within the standard model, seemingly immune to physicists' best tricks for neutralising infinity's effects. In extreme circumstances such as in a black hole's belly, Einstein's equations of general relativity, which describe gravity's workings, break down as matter becomes infinitely dense and hot, and space-time infinitely warped.

But it is at the big bang that infinity wreaks the most havoc. According to the theory of cosmic inflation, the universe underwent a burst of rapid expansion in its first fraction of a second. Inflation explains essential features of the universe, including the existence of stars and galaxies.

But it cannot be stopped. It continues inflating other bits of space-time long after our universe has settled down, creating an infinite "multiverse" in an eternal stream of big bangs. In an infinite multiverse, everything that can happen will happen an infinite number of times. Such a cosmology predicts everything – which is to say, nothing.

This disaster is known as the measure problem, because most cosmologists believe it will be fixed with the right "probability measure" that would tell us how likely we ›

LORENZO LINTHOUT/GALLERY STOCK

are to end up in a particular sort of universe and so restore our predictive powers. Others think there is something more fundamental amiss. "Inflation is saying, hey, there's something totally screwed up with what we're doing," says cosmologist Max Tegmark of the Massachusetts Institute of Technology (MIT). "There's something very basic we've assumed that's just wrong."

For Tegmark, that something is infinity. Physicists treat space-time as an infinitely stretchable mathematical continuum; like the line of real numbers, it has no gaps. Abandon that assumption and the whole cosmic story changes. Inflation will stretch space-time only until it snaps. Inflation is then forced to end, leaving a large, but finite, multiverse. "All of our problems with inflation and the measure problem come immediately from our assumption of the infinite," says Tegmark. "It's the ultimate untested assumption."

Disruptive influence

There are also good reasons to think it is an unwarranted one. Studies of the quantum properties of black holes by Stephen Hawking and Jacob Bekenstein in the 1970s led to the development of the holographic principle, which makes the maximum amount of information that can fit into any volume of space-time proportional to roughly one quarter the area of its horizon. The largest number of informational bits a universe of our size can hold is about 10^{122}. If the universe is indeed governed by the holographic principle, there is simply not enough room for infinity.

Certainly we need nothing like that number of bits to record the outcome of experiments. David Wineland, a physicist at the National Institute of Standards and Technology in Boulder, Colorado, shared the 2012 Nobel prize in physics for the world's most accurate measuring device, an atomic clock that could measure increments of time out to 17 decimal places – a record that's since been extended a decimal place or two. The electron's anomalous magnetic moment, a measure of tiny quantum effects on the particle's spin, has been measured out to 14 decimal places. But even the best device will never measure with infinite accuracy, and that makes some physicists very itchy. "I don't think anyone likes infinity," says Raphael Bousso of the University of California at Berkeley. "It's not the outcome of any experiment."

But if infinity is such an essential part of mathematics, the language we use to describe the world, how can we hope to get rid of it? Wildberger has been trying to figure that out, spurred on by what he sees as infinity's disruptive influence on his own subject. "Modern mathematics has some serious logical weaknesses that are associated in one way or another with infinite sets or real numbers," he says.

For the past decade or so, he has been working on a new, infinity-free version of trigonometry and Euclidean geometry. In standard trigonometry, the infinite is ever-present. Angles are defined by reference to the

"Inflation is saying, hey, there's something totally screwed up with what we're doing. Some basic assumption is very wrong"

CEDRIC LEFEBVRE/GALLERY STOCK

Sometimes what seems to be infinite is simply very, very long

circumference of a circle and thus to an infinite string of digits, the irrational number π. Mathematical functions such as sines and cosines that relate angles to the ratios of two line lengths are defined by infinite numbers of terms and can usually be calculated only approximately. Wildberger's "rational geometry" aims to avoid these infinities, replacing angles, for example, with a "spread" defined not by reference to a circle, but as a rational output extracted from mathematical vectors representing two lines in space.

Doron Zeilberger of Rutgers University in New Jersey, thinks the work has potential. "Everything is made completely rational. It's a beautiful approach," he says.

Then again, Zeilberger himself subscribes to a view of infinity so radical that it would have even the pre-Cantor greats of mathematics stirring in their coffins. While Wildberger's work is concerned with doing away with actual infinity as a real object used in mathematical manipulations, Zeilberger wants to dispose of potential infinity as well. Forget everything you thought you knew about mathematics: there is a largest number. Start at 1 and just keep on counting and eventually you will hit a number you cannot exceed – a kind of speed of light for mathematics.

That raises a host of questions. How big is the biggest number? "It's so big you could never reach it," says Zeilberger. "We don't know what it is so we have to give it a name, a symbol. I call it N_0." What happens if you add 1 to it? Zeilberger's answer comes by analogy to a computer processor. Every computer has a largest integer number that it can handle: exceed it, and you will either get an "overflow error" or the processor will reset the number to zero. Zeilberger finds the second option more elegant. Enough of the number line, stretching infinitely far in both directions. "We can redo mathematics postulating that there is a biggest number and make it circular," he says.

Hugh Woodin is a set theorist at Harvard University who has done seminal work on the nature of infinity and its relationship to maths (see "Ultimate logic", page 38). He is sceptical. "[Zeilberger] could be correct, of course. But to me the view is a limiting view. Why take it unless one has strong evidence that it is correct?" For Woodin, the success of set theory with all its infinities is reason enough to defend the status quo.

So far, finitist mathematics has received most attention from computer scientists and robotics researchers, who work with finite forms of mathematics as a matter of course. Finite computer processors cannot actually deal with real numbers in their full infinite glory. They approximate them using floating-point arithmetic – a form of scientific notation that allows the computer to drop digits from a real number, and so save on memory without losing its overall scope.

The idea that our finite universe might work similarly has a history. Konrad Zuse, a German engineer and one of the pioneers of floating-point arithmetic, built the world's first programmable electronic computer in his parents' living room in 1938. Seeing that his own machine could solve differential equations (which ordinarily use infinitely small steps to calculate the evolution of a physical system) without recourse to the infinite, he was persuaded that continuous mathematics was just an approximation of a discrete and finite reality. In 1969, Zuse wrote a book called *Calculating Space* in which he argued that the universe itself is a digital computer – one with no room for infinity.

Tegmark for his part is intrigued by the fact that the calculations and simulations that physicists use to check a theory against the hard facts of the world can all be done on a finite computer. "That already shows that we don't need the infinite for anything we're doing," he says. "There's absolutely no evidence whatsoever that nature is doing it any differently, that nature needs to process an infinite amount of information."

> "There is absolutely no evidence whatsoever that nature needs to process an infinite amount of information"

Seth Lloyd, a physicist and quantum information expert also at MIT, counsels caution with such analogies between the cosmos and an ordinary, finite computer. "We have no evidence that the universe behaves as if it were a classical computer," he says. "And plenty of evidence that it behaves like a quantum computer."

At first glance, that would seem to be no problem for those wishing to banish infinity. Quantum physics was born when, at the turn of the 20th century, physicist Max Planck showed how to deal with another nonsensical infinity. Classical theories were indicating that the amount of energy emitted by a perfectly absorbing and radiating body should be infinite, which clearly was not the case. Planck solved the problem by suggesting that energy comes not as an infinitely divisible continuum, but in discrete chunks – quanta.

The difficulties start with Schrödinger's cat. When no one is watching, the famous quantum feline can be both dead and alive at the same time: it hovers in a "superposition" of multiple, mutually exclusive states that blend together continuously. Mathematically, this continuum can only be depicted using infinities. The same is true of a quantum computer's "qubits", which can perform vast numbers of mutually exclusive calculations simultaneously, just as long as no one is demanding an output. "If you really wanted to specify the full state of one qubit, it would require an infinite amount of information," says Lloyd.

Down the rabbit hole

Tegmark is unfazed. "When quantum mechanics was discovered, we realised that classical mechanics was just an approximation," he says. "I think another revolution is going to take place, and we'll see that continuous quantum mechanics is itself just an approximation to some deeper theory, which is totally finite."

Lloyd counters that we ought to work with what we have. "My feeling is, why don't we just accept what quantum mechanics is telling us, rather than imposing our prejudices on the universe? That never works," he says.

For physicists looking for a way forward, however, it is easy to see the appeal. If only we could banish infinity from the underlying mathematics, perhaps we might see the way to unify physics. For Tegmark's particular bugbear, the measure problem, we would be freed from the need to find an arbitrary probability measure to restore cosmology's predictive power. In a finite multiverse, we could just count the possibilities. If there really were a largest number then we would only have to count so high.

Woodin would rather separate the two issues of physical and mathematical infinities. "It may well be that physics is completely finite," he says. "But in that case, our conception of set theory represents the discovery of a truth that is somehow far beyond the physical universe."

Tegmark, on the other hand, thinks the mathematical and physical are inextricably linked – the further we plunge down the rabbit hole of physics to deeper levels of reality, the more things seem to be made purely of mathematics. For him, the fatal error message contained in the measure problem is saying that if we want to rid the physical universe of infinity, we must reboot mathematics, too. "It's telling us that things aren't just a little wrong, but terribly wrong." ■

Careless pork costs lives...

...and other medical myths

It's not just tabloid newspapers that misrepresent medical statistics for dramatic effect, warn **Marianne Freiberger** and **Rachel Thomas**

TYPE the word "cancer" into the website search engine of the *Daily Mail*, a UK tabloid newspaper, and a wealth of information is just a mouse click away. Some of the reports are calming, most alarming – and all come with figures to back them up. Women who use talcum powder are 40 per cent more likely to develop ovarian cancer, says research. Cancer survival rates in the UK are among the worst in Europe, according to a study. The incidence of bowel cancer among the under-30s has soared by 120 per cent in 10 years, astonishing figures show.

The figures might make us worry for our health, but somehow we feel the better for their existence. Numbers help us make sense of the world: if you can put a number on a problem, then its extent is known and its impact can be circumscribed.

Yet that sense of solid certainty is all too often illusory. Statistics can be slippery, easily misused or misinterpreted. Nowhere is that more true than in the field of human health.

That's because the benefits of a particular medical treatment are often not obvious. "There are very few miracle cures. Most treatments require careful science to determine if there is any benefit and how big the benefit is," says David Spiegelhalter, a biostatistician at the University of Cambridge. "Working out the effects of an environmental risk factor is even more tricky," he adds. Saying anything sensible about human health requires large, reproducible clinical trials, and the careful observation of diverse populations – all of which implies the use of statistical methods to extract workable conclusions from the data.

The British epidemiologist Austin Bradford Hill recognised this when, in 1946, he ran the first trials in which participants were randomly assigned to two groups, one of which received the treatment and one of which didn't. One of these trials tested the effectiveness of the antibiotic streptomycin to treat tuberculosis, a condition that Bradford Hill himself had developed while serving in the first world war. After just six months, the results were so clear that they led to streptomycin being adopted as the standard treatment. In 1950, together with Richard Doll, Bradford Hill used statistical methods to provide the first convincing evidence that smoking causes lung cancer.

Used well, statistics are a powerful tool. But caution is required. Sample size, the design of a study and even the definition of terms or the way a number is presented can all affect the value of the headline statistics we are offered. Generally, we are not privy to these details.

What's more, decisions concerning health are often made at times of intense emotional stress. "People are very much influenced by culture, emotions and values when making judgements, and that's fine, that is part of being human," says Spiegelhalter. But it makes us all the more susceptible to seemingly incontrovertible numerical truths distilled into media headlines – and to the enthusiastic but sometimes equally misplaced insistence by researchers, doctors or advocates of a new treatment that it will do us good.

So when confronted with medical statistics, how do we know whether they are the real deal, or distorted before they get to us? How do errors creep in? What are the questions we need to ask to avoid falling for them?

YOUR NUMBER'S UP

Ratio bias

What would worry you more: being told that cancer kills 25 people out of 100, or that it kills 250 people out of 1000? Dumb question, you might say; both statements mean that a quarter of people die of cancer.

Yet such differences do matter – not to the risk itself, but to our perception of it. Those wishing to play up or play down a risk, whether to sell newspapers or a medical treatment, can follow the simple rule of "ratio bias". The bigger the number, the riskier the risk appears.

In one study of this effect, people rated cancer as riskier when told that it "kills 1286 people out of 10,000" than when told it "kills 24.14 people out of 100", even though the second statement equates to almost double the risk. Similarly, another study showed that 100 people dying from a particular form of cancer every day can be perceived as a lesser risk than 36,500 dying from the same disease each year, although the two are equivalent statements.

So when confronted with questions of risk, look carefully at the way the numbers are presented. And if you are comparing risks, make sure they are divided by the same number.

GETTY

MORE HARM THAN GOOD?

Relative versus absolute risk

Is there anything that has not been claimed to cause cancer? Over the years we have learned, among other things, that drinking very hot cups of tea leads to an eightfold increase in the risk of developing oesophageal cancer; that a quarter of a grapefruit a day increases breast cancer risk by 30 per cent in post-menopausal women; and that a daily bacon sandwich raises the likelihood of bowel cancer by 20 per cent. This last finding was encapsulated by UK tabloid *The Sun* in the headline "Careless pork costs lives".

These assertions may or may not be valid, but hidden within them is a more important and insidious source of confusion. The figures quoted measure relative risks: how much more likely you are to get ill when indulging in the supposedly dangerous substance or activity compared with not indulging. But they tell you nothing about what that increase in risk amounts to in absolute terms, so there is no way of telling whether it is something worth being concerned about.

"For an average person, the chance of getting bowel cancer at some point in their life is around 5 per cent," says Spiegelhalter. So a 20 per cent relative increase in bowel cancer risk translates to an absolute increase in risk from 5 per cent to 6 per cent – just 1 per cent. That's big enough not to ignore, but less of a deterrent to those who like their daily bacon sandwich.

Journalists are by no means the only ones who exploit the greater headline-grabbing potential of relative risk; health professionals do it too. "One of the most misleading, but rather common, tricks is to use relative risks when talking about the benefits of a treatment, while potential harms are given in absolute risks," says Spiegelhalter.

This technique is known as mismatched framing. In his book *Reckoning with Risk*, psychologist Gerd Gigerenzer of the Max Planck Institute for Human Development in Berlin, Germany, quotes the example of a patient information leaflet concerning hormone replacement therapy. It claimed that HRT cuts the risk of bowel cancer by 50 per cent (a relative risk), but leads to 6 extra cases of breast cancer per 1000 women (an absolute risk). At first glance, the benefit here seems to hugely outweigh the additional breast cancer risk of just 0.6 per cent.

But until we know the absolute rates of bowel cancer in the target population, we are none the wiser. Assuming that rate is 5 per cent, as it is in the general population, the reduction in risk is 2.5 per cent, putting the benefit to harm ratio in a very different light.

Once you are aware of this trick, it's relatively easy to spot, but this doesn't eradicate it even from peer-reviewed medical journals. According to a study published in 2007, one-third of papers reporting on the benefits and harms of medical interventions in the *BMJ*, *The Lancet* and *The Journal of the American Medical Association* presented them using a mixture of different measures.

EVGENIY IVANOV/GETTY

Some like it hot – but is tea that's too hot a significant cancer risk?

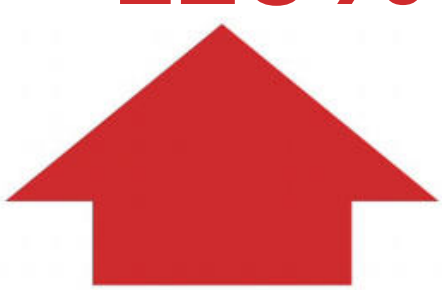
Scary or not?

Different ways of presenting the same data can greatly influence our perception of risk

"Bowel cancer soars by 120% among the under 30s"

Daily Mail, 31 March 2009

+120%



137 people under 30 were diagnosed with bowel cancer in 2006 in England & Wales

up from 63 cases in 1997

"One-third of academic papers reporting medical benefits and harms used a mixture of statistical measures"

"Two things moving in tandem does not necessarily mean that one thing is causing the other to move"

TV KILLS

Correlation vs causation

It isn't surprising that a study with the title "Television viewing time and mortality" grabbed the headlines. It asked 8800 people about their health, lifestyle and television watching behaviour, and then followed them over the next six years, during which time 284 of them died. Among people who spent more than 4 hours a day in front of the TV, it found, the risk of their dying within the period of the study was 46 per cent higher than among those who watched less than 2 hours a day.

The sort of headlines generated – "TV kills, claim scientists" – were also predictable. But this is one case where two variables moving in tandem (correlation, in other words) does not necessarily mean that the change in one is responsible for change in the other (causation). In fact, the researchers were not primarily interested in TV viewing. They wanted to measure the amount of time people spent sitting still, and used TV watching as a shorthand for this; they explicitly excluded time spent watching TV while doing other, active things, such as ironing.

"At best, this study shows that sedentary behaviour, for which hours of TV watching is a proxy, is associated with modest elevations in death from heart disease and from all causes," says Nigel Hawkes, a health journalist and formerly the director of Straight Statistics (straightstatistics.org), a campaign to improve the use of statistics in the public arena. "There is nothing intrinsic in television that makes people more likely to die."

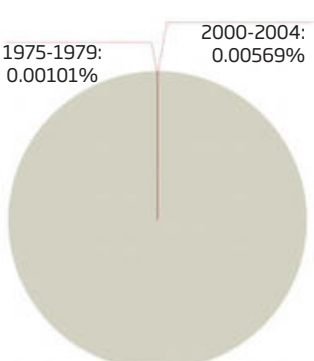
You don't have to look far to find confounding variables that might have been at work. People with certain underlying health problems sit or lie still for long periods, possibly in front of the TV, and these problems might also be associated with a raised risk of early death. Despite the study's apparent conclusions, it's probably still safe to switch on and zone out.

Before assigning cause and effect, it is essential to read between the lines. Bradford Hill identified the crucial question: Is there any other way of explaining the set of facts before us; and is any such explanation equally, or more, likely than cause and effect? The answer needs to be a resounding no.

Fox News
27 August 2008

"Throat cancer among white men up 400% in 30 years"

Percentage of white male population in the US diagnosed annually with adenocarcinoma of the oesophagus
Journal of National Cancer Institute, vol 100, p 1184



SIZE MATTERS

Clinical trial design

"Over 80 per cent of women say that this shampoo leaves their hair healthier and shinier." Such claims are common in advertising for all manner of consumer products. What they might not tell you is that only five women tested the shampoo. And of the four who certified its miraculous effect, one or two probably ended up with nicer hair purely by chance, or simply imagined the results.

Similar caveats apply to the effectiveness of medical treatments. Curing six out of 10 patients is promising. Curing 300 out of 500 is the same success rate, but far more convincing. "The sample size in a test is absolutely crucial in deciding whether any apparent improvement could have happened by chance alone," says Spiegelhalter.

The standard procedure for such trials is the one established by Bradford Hill more than 60 years ago: new medical treatments are tested in randomised controlled trials (RCTs), in which volunteers are randomly allocated to a study group that receives the new treatment or a control group that receives a placebo or existing treatment. "You can think of an RCT almost as a measuring instrument to measure a treatment's effectiveness," says Sheila Bird of the UK Medical Research Council Biostatistics Unit in Cambridge. To make sure any instrument is sensitive enough for its job, you need to assess how big an effect it is expected to measure.

Working out the size of the expected effect requires an analysis of past studies or the results of tests on animals. In the case of an RCT, the smaller the expected effect, the more people you need to enrol in your trial, and vice versa.

Another important consideration is the level of significance the trial is expected to achieve – that is, the likelihood that a useless treatment will register the effect you are after as a result of chance alone. RCTs are usually designed to achieve a 5 per cent significance level. This means that even if the drug is useless, it will register a positive result by chance in one out of 20 trials. For that reason, says Spiegelhalter, drug licensing authorities do not usually consider a single study sufficient evidence to approve a new drug. Repeat trials are needed.

So next time you hear of public acclaim for a miracle cure or wonder shampoo, ask three questions. How many people was it tested on? Was it tested in an RCT? And was the result confirmed by a second, independent test?

"Rudy Giuliani claimed you were only half as likely to survive prostate cancer in the UK as in the US. He was right – but also wrong"

Is the secret to a long life laying off the grapefruit?

EDWARD POND/GETTY

DIE ANOTHER DAY

Survival vs mortality

There can be few things in US politics more poisonous than discussions about healthcare. Over the years, the arguments have been accompanied by all sorts of dodgy claims and counterclaims, often with statistical evidence to back them up.

Take the statement by former New York City mayor Rudy Giuliani in his campaign to win the 2008 Republican presidential nomination. He quoted the chance of a man surviving prostate cancer – a disease he had himself experienced – as 82 per cent in the US, and compared this with a chance of just 44 per cent under the UK's taxpayer-funded National Health Service.

Survival rates a factor of two apart in two comparably developed countries? If right, surely that would be a damning indictment of the deadly inadequacy of socialised medicine. And there's no doubting Giuliani's figures were right.

Right – but also misleading. "Giuliani's numbers are meaningless for making comparisons across groups that differ dramatically in how the diagnosis is made," observed Gigerenzer and colleagues in a 2008 paper on risk communication.

That is because Giuliani was quoting five-year survival rates – the number of people diagnosed with a disease in a given year who are still alive five years later. But while prostate cancer in the US is generally diagnosed through screening, in the UK it is diagnosed on the basis of symptoms. Screening tends to pick up the disease earlier, leading to one source of bias in the comparison.

Suppose that of a group of men with prostate cancer all die at the age of 70. If the men do not develop symptoms until they are 67 or later, the five-year survival rate based on a symptoms approach is 0 per cent. Suppose, instead, that screening had picked up the cancer in all of these men at age 64. The five-year survival rate in this case is 100 per cent, despite the fact that mortality is the same. Better survival rates don't necessarily indicate a better outcome.

That is obviously an oversimplification, as earlier diagnosis through screening presumably increases the chance that corrective measures can be taken. But screening is not 100 per cent accurate. First there are false positives, in which the test incorrectly flags a healthy person as having cancer. Prostate screening also picks up non-progressive cancers, which will never lead to symptoms, let alone death. The exact extent of this overdiagnosis is unclear, but a rough estimate is that 48 per cent of men diagnosed in this way don't have a progressive form of the cancer.

Tricky comparison

False diagnosis and overdiagnosis both result in unnecessary treatment, and, potentially, significant harm – in the case of prostate cancer, men left impotent and incontinent. But overdiagnosis also inflates the five-year survival rate by including men who would not have died of prostate cancer anyway. "In the context of screening, survival is a biased metric," says Gigerenzer. "The bottom line is that to learn which country is doing better, you need to compare mortality rates."

The annual mortality from a disease is the proportion of people in the whole population who die from it in a given year. So which comes out better, the US or the UK? Figures from the period 2003 to 2007 published by the US National Cancer Institute indicate an age-adjusted mortality from prostate cancer of 24.7 per 100,000. Similar figures from Cancer Research UK for 2008 point to a mortality of 23.9 per 100,000. In statistical terms, that is a dead heat. Higher survival does not necessarily mean fewer deaths.

This kind of bias makes it tricky to compare survival rates in different countries, a difficulty often explicitly acknowledged by the authors of academic studies that use the metric. Equally often, that subtlety is overlooked by politicians and journalists in search of a shocking sound bite or headline.

So next time you are told that one country outperforms or underperforms another on some vital metric of health, take a close look at whether it is survival or mortality that is being quoted. If it's the former, take the figure with a pinch of salt. Be aware, though, that this may increase your risk of heart disease. ■

NEED TO KNOW

HOW TO PLAY THE GAME

Game theory is the science of strategic thinking – grasp it to win at life

IN THE film *A Beautiful Mind*, mathematician John Nash and his Princeton grad student buddies are sitting in a smoky bar when a group of women walk in. As the men tease each other about their chances, Nash is struck with inspiration. Is there a logical, mathematical way of working out the best strategy for each man getting a date? Next thing you know he's shambling out of the bar, and spends the night furiously scribbling equations.

In a ham-fisted, Hollywood sort of way, this imagined episode does hint at how game theory, the branch of maths Nash helped to make famous, can apply to our everyday lives. In fact, we use it all the time without even realising. "Every time you think about what you should do in terms of what someone else will do in response, you're doing rudimentary game theory," says Kevin Zollman of Carnegie Mellon University in Pittsburgh, Pennsylvania.

When most of us need to think through situations several steps ahead or when they involve more than just a few people, we make mistakes. But delve a little into the theory, and you can make smarter moves in your own life.

Lesson one is that there are different sorts of games. Broadly speaking, there are zero-sum games, in which one player gains what the other loses, and variable-sum games, in which players have both common and opposed interests.

An example of a zero-sum game would be chess or poker. When you win, your opponent automatically loses and vice-versa. These sorts of situations don't crop up much in everyday life. Variable-sum games are more common and more complex. They are exemplified by what's known as the prisoner's dilemma, a scenario in which the punishment you receive for a crime depends on both your plea and that of an accomplice. You don't know how your accomplice will behave, but game theory organises the possible outcomes into a pay-off matrix that allows you to think through the various possible outcomes (see "Decisions in the frame", below left).

> "One way to win at chicken is to throw the steering wheel out of the window"

Credible deterrence

It turns out, perhaps counter-intuitively, that your best option is for both you and your accomplice to confess. This decision is what's known as a Nash equilibrium because neither party can benefit from making a different choice while the other party's choice stays the same.

Nash died in 2015, but his contribution to game theory, including the equilibrium idea, helped him win a share of the 1994 Nobel prize in economics. Lessons from the discipline have been applied all over the place, from politics and diplomacy to economics and business. It helped the US formulate its nuclear deterrent strategy during the cold war, for instance. Today, broadcasters use it to jostle for the rights to air top-level sports fixtures.

But individuals can harness insights from game theory, too. One example is understanding the power of "credible commitment", says Rakesh Vohra, an economist at the University of Pennsylvania in Philadelphia. This concept is best described by a game of chicken. Think of two cars accelerating towards each other; the loser is the one who swerves. Here, the Nash equilibria are the two situations in which one player swerves and not the other.

But a game theory analysis shows one of the drivers can force an outcome by changing the game's rules – for example by removing the steering wheel and throwing it out of the window, so the other driver must swerve to avoid destruction. "You're making your opponent recognise that you have no choice but to take a particular action, which then forces them to do what you want them to do," says Vohra. "Limiting your options can sometimes make you better off."

The same principle can be applied to buying rather than crashing a car. Do your research on prices and make a take-it-or-leave-it offer. "By committing yourself, you force the seller to make a choice: either sell at that price or make no sale at all," says Vohra. This reasoning applies to any situation in which two competing parties have to negotiate a price: agreeing a salary for a new job, for instance.

Even the greatest game theorists won't always get it right, however. Game theory assumes we act rationally all the time – and we don't. Even the experts will sometimes be thrown off by the quirks of human behaviour.

Daniel Cossins

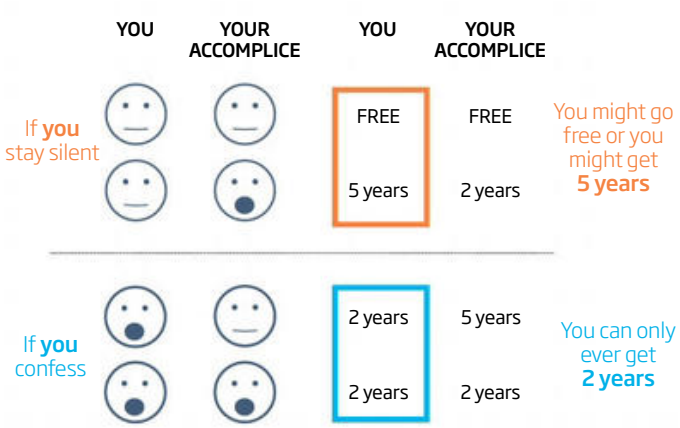
Decisions in the frame

Game theory provides us with a framework to make decisions when we have incomplete information – such as in the famous prisoner's dilemma

You and an accomplice are being held separately for a crime. The maximum penalty is 5 years – or 2 if you confess

The police can only prosecute if one of you confesses – if neither does, you walk free

Should you stay silent or should you confess?



NEED TO KNOW

HOW NUMBERS GROW

The law of exponential growth can make or break you

THE greatest shortcoming of the human race is our inability to understand the exponential function." These are the words of the late Albert Bartlett, a physicist at the University of Colorado, Boulder, whose lectures on the subject became a YouTube hit.

Take saving for retirement. "Start early" is the mantra, but it is easy to overlook just how much difference a few years can make. It all comes down to exponential growth – an often abused term that refers to anything that grows in proportion to its current value. It dictates that a forward-thinking 18-year-old can retire as a millionaire at 65 by investing around £250 a month with an annual return of 7 per cent.

That figure might sound high by today's standards, but it's a rough average of the stock market return since 1960. The surprise is that when our saver reaches 55, the savings will amount to a little under £500,000. Thanks to the power of compound interest, however – exponential growth by another name – it will double to £1 million just 10 years later. Wait until you are 30 to start saving that £250 and you will only reach about half as much (see diagram, below). Starting at 30, you would actually need to save more than £600 a month to make it to £1 million by 65.

Start 'em young

A short delay in saving for your pension can have a huge impact on how much you accrue

£1.134 M
Saving £250 a month starting at 18

£478 K
Saving £250 a month starting at 30

Exponential growth's stealth factor is nicely illustrated by the story of the man who invented chaturanga, an Indian precursor to chess. He presented his king with a beautifully laid out board divided into 64 squares and when asked to name his reward, requested a grain of wheat to be placed on the first square, two on the next, four on the third, and so on.

It sounded a modest reward, but had the king obliged across the board, he would have given away more than 18 billion billion grains. Fail to understand exponential growth, and our debts can rapidly spiral out of control too. This is an engine for creation and destruction wrapped up in deceptively simple maths. In reference to the chaturanga legend, US futurist Ray Kurzweil refers to the sudden changes that spring from exponential growth as the "second half of the chessboard".

The number of transistors that can fit on a electronic chip provides an example. Over the past few decades, it has roughly doubled every 18 months, a phenomenon known as Moore's law. The accelerating effect of exponential growth explains why we spent 25 years with bulky desktop computers before rapidly switching over to smartphones. Kurzweil is famed for believing that this sort of technological growth will lead to an event called the singularity, when computers will become powerful and smart enough to improve themselves and outpace us all.

The spread of viruses often works in a similar way: one ill person infects a few others, who in turn each infect a few more, until we've got an epidemic on our hands. Immunisation acts as the limiting factor, which is why the world scrambled to treat last year's Ebola outbreak, which at one point saw the number of known cases doubling every few weeks.

When it comes to exponential growth, you can't trust your short-term instincts. Whether it's finance or technology, the largest changes won't happen for some time. But when they do, your whole world can be turned upside down in an instant. **Jacob Aron**

HOW TO THINK ABOUT PROBABILITY

THE MONTY HALL PROBLEM

Suppose you're on a game show,
and you're given the choice of three doors

"Pick a door" 1 2 3

Behind one door is a car; behind the others, goats

You pick a door, say **No. 1**, and the host, who knows what's behind the doors, opens another door to reveal a goat

The host then says to you,

"Do you want to switch to number 2?"

Counter-intuitively, you should **switch**. Here's why

YOU PICK 1	2	HOST OPENS	3	You STICK	You SWITCH
Car	Goat	or	Goat	✓	✗
Goat	Car		Goat	✗	✓
Goat	Goat		Car	✗	✓
The fact that the host knows what is behind the doors affects your chances so the win ratios are				*1/3*	*2/3*

The same applies if you pick 2 or 3

PROBABILITY is one of those things we all get wrong... deeply wrong. The good news is we're not the only ones, says John Haigh, formerly of the University of Sussex in Brighton, UK. "Many pure mathematicians claim that probability has many unreasonable answers."

Take the classic problem of a class of 25 schoolchildren. How likely is it that two of them share the same birthday? The common-sense answer is that it is not implausible, but quite unlikely. Wrong: it's actually just under 57 per cent.

Or the celebrated Monty Hall problem, named after the former host of US television game show *Let's Make a Deal*. You're playing a game in which there are three doors, one hiding a car, two of them goats (see diagram, left). You choose one door; the host of the game then opens another, revealing a goat. Assuming you'd rather win a car than a goat, should you stick with your choice or swap?

The naive answer is that it doesn't matter: you now have a 50-50 chance of striking lucky with your original door. Wrong again.

But if probability makes even experts grumble, how do we get it right? Simple, says mathematician Ian Stewart of the University of Warwick in the UK: do things the hard way. "The important thing with probability is not to intuit it," he says. Think carefully about how the problem is posed and do your sums diligently, and you'll arrive at the right answer – eventually.

With the birthday problem, the starting point is to realise that you're not interested in individual schoolchildren, but pairs. In a class of 25, there are 300 pairs to consider and, in most years, 365 days on which each might share a birthday.

Factor all that in, and you end up crunching some truly astronomical numbers to arrive at the answer. "Any coincidence like that is remarkable in itself, but when you ask how many times it would happen, that number is so vast it's not remarkable at all," says Haigh.

With the Monty Hall problem, meanwhile, the chance you chose the right door in the first place is 1/3 – and that doesn't change whatever happens afterwards. Since the host has revealed a goat, there is now a 2/3 probability that the car is behind the other door – and you are better off swapping.

There are a few caveats: if the host is so devious as only to open a door if you chose the right one in the first place, you'd be mad to swap. Ditto if you want a goat rather than the car. That illustrates another important rule in thinking about probability, says Haigh. "It is very important to know your assumptions. Very subtle changes can change the outcome."

All this is very well when the boundaries of the problem are clear and the possible outcomes quantifiable. Toss a fair coin and you know you have a 50 per cent chance of heads – because you can repeat the exercise over and over again if necessary.

But what about a 50 per cent chance of rain today, or of a horse with even odds winning a race? No amount of expert advice can help us assess the true worth of such "subjective" probabilities, which are fluid and often based on inscrutable expertise or complex modelling of an unpredictable world. Sometimes you do just have to go with your gut instinct – and be prepared to be wrong. Richard Webb

EUGENIA LOLI

PROBABILITY WARS

Can't get your head around uncertainty? You're in good company, says **Regina Nuzzo**

WE ARE in a bar, and agree to toss a coin for the next round. Heads, I pay; tails, the drinks are on you. What are your chances of a free pint?

Most people – sober ones, at least – would agree: evens.

Then I flip the coin and catch it, but hide in it the palm of my hand. What's your probability of free beer now?

Broadly speaking, there are two answers: (1) it is still 50 per cent, until you have reason to think otherwise; (2) assigning a probability to an event that has already happened is nonsense.

Which answer you incline towards reveals where you stand in a 250-year-old, sometimes strangely vicious debate on the nature of probability and statistics. It is the spat between frequentist and Bayesian statistics, and it is more than an esoteric problem. "The frequentist-Bayesian debate is the only scientific controversy that actually does affect everybody's life," says Larry Wasserman of Carnegie Mellon University in Pittsburgh, Pennsylvania. A drugs company testing a new drug can come to apparently very different conclusions according to which method it uses to analyse its results. A jury might reach a different decision after hearing evidence presented in frequentist and Bayesian terms. "It's not just philosophy, and it's not just mathematics. It really is concrete," says Wasserman.

The two approaches have often seemed at loggerheads. But statisticians are slowly coming to a new appreciation: in a world of messy, incomplete information, the best way might be to combine the two very different worlds of probability – or at least mix them up a little.

To fully appreciate the profundity of our bar bet, let's start with an old T-shirt slogan: "Statistics means never having to say you're certain." Drawing conclusions without all the facts is the bread and butter of statistics. How many people in a country support legalising cannabis? You can't ask all of them. Is a run of hotter summers consistent with natural variability, or a trend? There's no way to look into the future to say definitively.

LIFETIME CHANCE OF BEING KILLED BY A DOG IN THE US: 1 IN 103,798
Source: US National Safety Council

Answers to such questions generally come with a probability attached. But that single number often masks a crucial distinction between two different sorts of uncertainty: stuff we don't know, and stuff we can't know.

Can't-know uncertainty results from real-world processes whose outcomes appear random to all who look at them: how a die rolls, where a roulette wheel stops, when exactly an atom in a radioactive sample will decay. This is the world of frequentist probability, because if you roll enough dice or observe enough atoms decaying, you can get a reasonable handle on the relative frequency of different outcomes, and can construct a measure of their probabilities.

Ignorance is Bayesian

Don't-know uncertainty is more slippery. Here individual ignorance, not universal randomness, is at play. What's the sex of an unborn child? We don't yet know – although it is already a given. What horse will win a race that hasn't yet started? That is not a given, but studying previous form might give us a better idea than we would otherwise have had.

How to approach these different types of uncertainty divides frequentists and Bayesians. A strict frequentist has no truck with don't-know uncertainty, or any probability measure that can't be derived from repeatable experiments, random number generators, surveys of a random population sample and the like. A Bayesian, meanwhile, doesn't bat an eyelid at using other "priors" – knowledge gleaned from the form book in a horse race, for example – to fill in the gaps (see "The bookie", right). "Bayesians are happy to put probabilities on statements about the world," says Tony O'Hagan, a statistician at the University of Sheffield, UK, who researches Bayesian methods. "Frequentists aren't."

The coin-in-the-pub example shows where these two views diverge. Before I flip the coin, frequentist and Bayesian probabilities line up: 50 per cent. Then the source of uncertainty changes from intrinsic randomness to personal ignorance. Only if you were inclined to Bayesian ways of working would you be happy to quote a probability figure. That figure might be 50 per cent – or perhaps a telltale flicker of a victorious smile on my face might persuade you to downgrade your chances of a free drink to just 20 per cent, say. "In the Bayesian approach we try to answer questions by bringing all the relevant evidence to bear on it, even when the contribution of some of that evidence to the question depends on subjective judgements," says O'Hagan.

Bayesianism takes its name from the English mathematician and Presbyterian minister Thomas Bayes. In an essay published in 1763, two years after his death, he set out a new approach to a fundamental puzzle: how to work backwards from observations to hidden causes when your information is incomplete. Imagine you have a box of a dozen doughnuts, half cream, half jelly-filled. It's relatively straightforward to calculate the probability of pulling out five jelly doughnuts in a row. But the backwards problem, working out the probable contents of an unknown box when you've just pulled out five jellies, is trickier. Bayes's innovation was to provide the seed of a mathematical framework that allowed you to start with a guess (perhaps you've bought boxes of doughnuts from that store before), and refine it as further data came to light.

In the late 18th and early 19th centuries, Bayesian-style methods helped tame a range of inscrutable problems, from estimating the mass of Jupiter to calculating the number of boys born worldwide for every girl. But it gradually fell out of favour, victim of a dawning era of big data. Everything from improved astronomical observations to newly published statistical tables of mortality, disease and crime conveyed a reassuring air of objectivity. Bayes's methods of educated guesswork seem hopelessly old-fashioned, and rather unscientific by contrast. Frequentism, with its emphasis on dispassionate number crunching of the results of randomised experiments, came increasingly into vogue.

The advent of quantum theory in the early 20th century, which re-expressed even reality in the language of frequentist probability (see "Random reality", page 106), provided a further spur to that development. The two strands of thought in statistics gradually drifted further apart. Adherents ended up submitting work to their own journals, attending ›

THE BOOKIE
ALAN GLYNN
Head of Sports Trading at bookmakers Paddy Power

How do you calculate odds?
You analyse the teams or competitors with the stats you have available. For a Premier League football game, for example, you look at which team is better in the long run, which one has played better recently, where the game is being played, which players are available and any other factors such as how important this particular game is to each team. Having weighed up all these things, you come up with a probability that each team might win.

How much is down to maths and how much to human judgement?
We use algorithms and mathematical models, but very few bookmakers would generate prices and be confident of them using a computer program alone. You need human intervention to make sure everything has been taken into account. For the most part, the final figure comes from the trader's head. This is what bookmaker traders do every day. They're good at it.

There are basic rules to follow. For example, in the Premier League, about 44 per cent of games are won by the home team, 26 per cent by the away team and 30 per cent are a draw. If you had two teams that looked as good as each other, then those are the percentages you would put on the match. That is your starting point. If you thought the home team was slightly better, you might give them 46 per cent chance of a win instead of 44 per cent.

How are the odds affected by the way punters bet?
Usually not at all. In something like a Premier League game, where the public has as much information as we do, we'd be very confident in the odds we set and wouldn't change our prices readily. That's not to say we would never pay attention to where the money goes. You can't know every last detail about every event. If you have the 500th tennis player in the world playing the 550th, clearly information is at a premium. If we saw a run of bets on a game like that we would definitely change our odds, because it could be a key indicator of what's going to happen.

their own conferences and even forming their own university departments. Emotions often ran high. The author Sharon Bertsch McGrayne recalls that when she started researching her book on the history of Bayesian ideas, *The Theory That Would Not Die*, one frequentist-leaning statistician berated her down the phone for attempting to legitimise Bayesianism. In return, Bayesians developed a sort of persecution complex, says Robert Kass at Carnegie Mellon. "Some Bayesians got very self-righteous, with a kind of religious zealotry."

BRUCE FORSTER/GETTY

Flexible friend

In truth, though, both methods have their strengths and weaknesses. Where data points are scant and there is little chance of repeating an experiment, Bayesian methods can excel in squeezing out information. Take astrophysics as an example. A supernova explosion in a nearby galaxy, the Large Magellanic Cloud, seen in 1987, provided a chance to test long-held theories about the flux of neutrinos from such an event – but detectors picked up only 24 of these slippery particles. Without data, frequentist methods fell down – but the flexible, information-borrowing Bayesian approach provided an ideal way to assess the merits of different competing theories.

It helped that well-grounded theories provided good priors to start that analysis. Where these don't exist, a Bayesian analysis can easily be a case of garbage in, garbage out. It's one reason why courts of law have been wary of adopting Bayesian methods, even though on the face of it they are an ideal way to synthesise messy evidence from many sources (see "Justice you can count on", page 60). In a 1993 New Jersey paternity case that used Bayesian statistics, the court decided jurors should use their own individual priors for the likelihood of the defendant having fathered the child, even though this would give each juror a different final statistical estimate of guilt. "There's no such thing as a right or wrong Bayesian answer," says Wasserman. "It's very postmodern."

Finding good priors can also demand an impossible depth of knowledge. Researchers searching for a cause for Alzheimer's disease, for instance, might test 5000 genes. Bayesian methods would mean providing 5000 priors for the likely contribution of each gene, plus another 25 million if they wanted to look for pairs of genes working together. "No one could construct a reasonable prior for such a high dimensional problem," says Wasserman. "And even if they did, no one else would believe it."

To be fair, without any background information, standard frequentist methods of sifting through many tiny genetic effects would have a hard time letting the truly important genes and combinations of genes rise to the top of the pile. But this is perhaps a problem more easily dealt with than conjuring up 25 million coherent Bayesian guesses.

Frequentism in general works well where plentiful data should speak in the most objective way possible. One high-profile example is the search for the Higgs boson, completed in 2012 at the CERN particle physics laboratory near Geneva, Switzerland. The analysis teams concluded that if in fact there were no Higgs boson, then a pattern of data as surprising as, or more surprising than, what was observed would be expected in only one in 3.5 million hypothetical repeated trials. That is so unlikely that the team felt comfortable rejecting the idea of a universe without a Higgs boson.

That wording may seem convoluted, and highlights frequentism's main weakness: the way it ties itself in knots through its disdain for all don't-know uncertainties. The Higgs boson either exists or it doesn't, and any inability to say one way or the other is purely down to lack of information. A strict frequentist can't actually make a direct statement of the probability of its existing or not – as indeed the CERN researchers were careful not to (although certain sections of the media and others were freer).

"WHERE THERE'S NO WELL-GROUNDED THEORY, BAYESIAN STATISTICS CAN BE GARBAGE IN, GARBAGE OUT"

Head-to-head comparisons can point to the confusions that can arise, as was the case with a controversial clinical trial of two heart-attack drugs, streptokinase and tissue plasminogen activator, in the 1990s. The first, frequentist analysis gave a "p value" of 0.001 to a study seeming to show that more patients survived after the newer, more expensive tissue

"USING THE TWO TYPES OF PROBABILITY TOGETHER CAN TRUMP EITHER ALONE"

plasminogen activator therapy. This equates to saying that if the two drugs had the same mortality rate, then data at least as extreme as the observed rates would occur only once in every 1000 repeated trials.

That doesn't mean the researchers were 99.9 per cent certain the new drug was superior – although again it is often interpreted that way. When other researchers conducted a Bayesian reanalysis using the results of previous clinical trials as a prior, they found a direct probability of the new drug being superior of only about 17 per cent. "In Bayesianism we're directly addressing the question of interest, talking about how likely it is to be true," says David Spiegelhalter of the University of Cambridge. "Who wouldn't want to talk about that?"

Perhaps it's just a case of horses for courses, but don't the strengths and weaknesses of the different approaches suggest we might be better off combining elements of both? Kass is one of a new breed of statisticians doing just that. "To me statistics is like a language," he says. "You can be conversant in both French and English and switch back and forth comfortably."

Stephen Senn, a drugs statistician at the Luxembourg Institute of Health agrees. "I use what I call 'mongrel statistics', a little bit of everything," he says. "I often work in a frequentist mode, but I reserve the right to do Bayesian analyses and think in a Bayesian way."

Kass points to an analysis he and his colleagues did on the firing rates of a couple of hundred neurons in the visual-motor region of the brain in monkeys. Prior work in basic neurobiology provided them with information on how fast these neurons should be firing, and how quickly the r ate might change over time. They incorporated this into a Bayesian approach, then switched gears to evaluate their results under a standard frequentist framework. The Bayesian prior gave the methods enough of a kick-start to allow frequentist methods to detect even tiny differences in a sea of noisy data. The two approaches together trumped either method alone.

Sometimes, Bayesian and frequentist ideas can be blended so much they create something new. In large genomics studies, a Bayesian analysis might exploit the fact that a study testing the effects of 2000 genes is almost like 2000 parallel experiments, and cross-fertilise the analyses, using the results from some to establish priors for others and using that to hone the conclusions of a frequentist analysis. "This approach gives quite a bit better results," says Jeff Leek of Johns Hopkins University in Baltimore, Maryland. "It's really transformed the way we analyse genomic data."

It breaks down barriers, too. "Is this approach frequentist? Bayesian?" asked Harvard University statistician Rafael Irizarry in a blog post. "To this applied statistician, it doesn't really matter."

Not that the arguments have entirely gone away. "Statistics is essentially the abstract language that science uses on top of data to tell stories about how nature works, and there is not one unique way to tell stories," says Kass. "Two hundred years from now there might be some breakthrough connecting Bayesianism and frequentism into a grand synthesis, but my guess is that there will always be at least one versus the other."

So in all probability, in two centuries' time two people will still be sitting on pub stools arguing about their chances of free beer. ■

THE GAMBLER
VANESSA SELBST

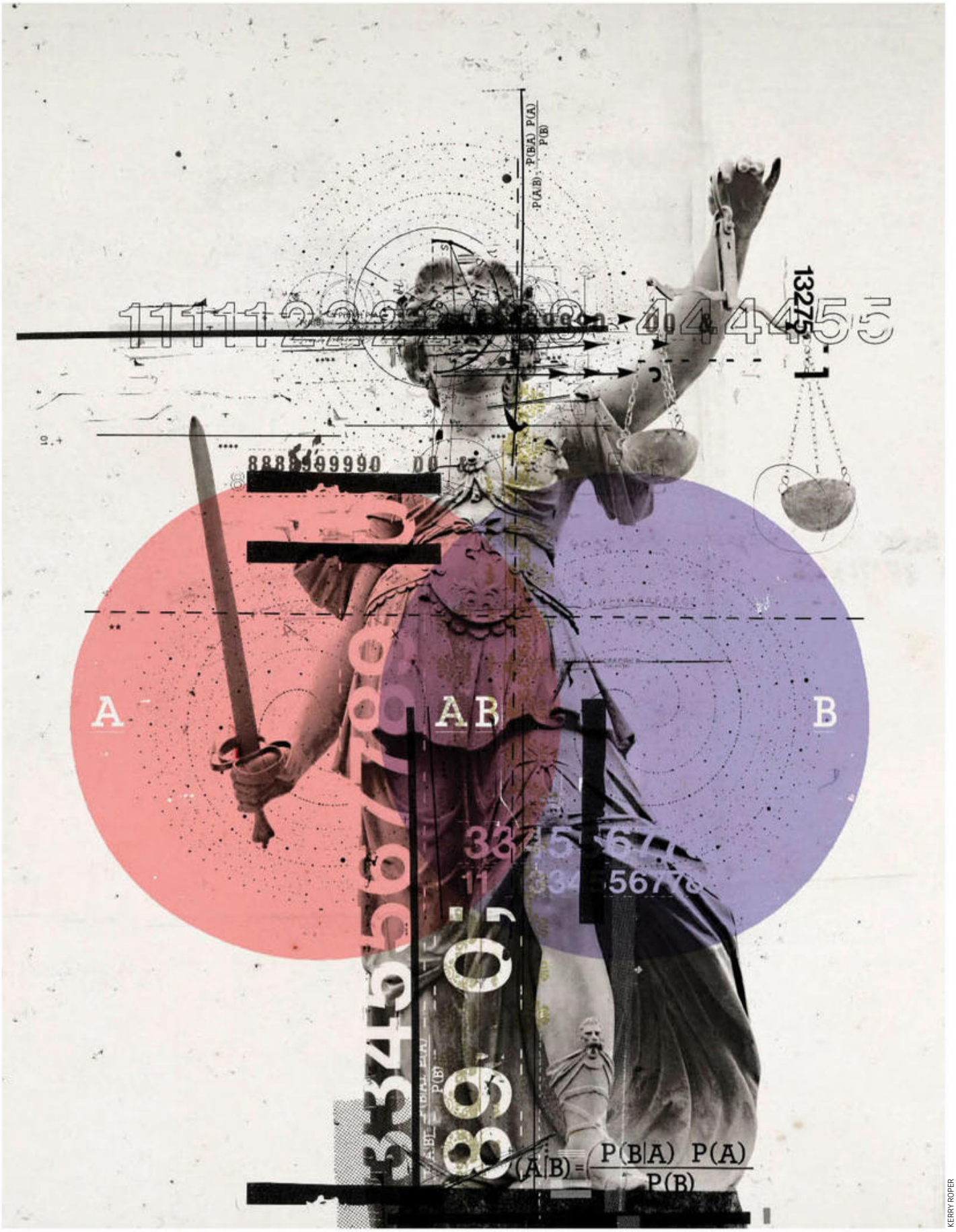
Highest earning female poker player of all time

So, that old question: how much is poker about luck?
Luck is a huge factor. Your job as a poker player is to identify the situations in which you have a very good chance of winning and risk as much money as possible. The skilful players will give themselves a better chance of winning, but no matter how good you are, there is so much luck involved in any specific hand. Of course, given the law of large numbers, the luck eventually runs out. So the more tournaments you play, the less luck is involved in the game.

How important is it to understand probability theory?
There is a lot of simple maths you need to know and memorise. For instance, what are the odds of making a flush [five cards of the same suit] if you have two in your hand and there are two on the table? After that, the maths is just one of many factors you can use to figure out what someone has in her hand. There are "maths players" who rely mostly on that aspect, but normally you use some combination of maths, deductive reasoning and psychology. For me, reasoning is the biggest part of it – taking all the possibilities and eliminating each possibility until I end up with the most likely scenario.

Do you think of yourself as a gambler?
I don't really like to gamble, which is a funny thing for a professional gambler to say. But I prefer not to. If I bet this hand and I know I have a 60 per cent chance of winning, I would much rather you paid me 60 per cent of the pot right now than allow the cards to determine the fate of the hand. But unfortunately for me, that's not part of the game.

How do you cope with losing?
You will inevitably have down swings – I'm in one right now. There have been lots of situations where I've had an 80 per cent chance of winning and lost. It's happened an incomprehensible number of times in a row, something like 20 of the last 25 tournaments. Those situations can be demoralising, but poker players have to be rational.



KERRY ROPER

Justice you can count on

A poor grasp of probability can lead to all sorts of legal problems, says Angela Saini

SHAMBLING sleuth Columbo always gets his man. Take the society photographer in a 1974 episode of the cult US television series who has killed his wife and disguised it as a bungled kidnapping. It is the perfect crime – until the hangdog detective hits on a cunning ruse to expose it. He induces the murderer to grab from a shelf of 12 cameras the exact one used to snap the victim before she was killed. "You just incriminated yourself, sir," says a watching police officer.

If only it were that simple. Killer or not, anyone would have a 1 in 12 chance of picking the same camera at random. That kind of evidence would never stand up in court.

Or would it? In fact, such probabilistic pitfalls are not limited to crime fiction. "Statistical errors happen astonishingly often," says Ray Hill, a mathematician at the University of Salford, UK, who has given evidence in several high-profile criminal cases. "I'm always finding examples that go unnoticed in evidence statements."

The root cause is a sloppiness in analysing odds that can sully justice and even land innocent people in jail. With ever more trials resting on the "certainties" of data such as DNA matches, the problem is becoming more acute. Some mathematicians are calling for the courts to take a crash course in the true significance of the evidence put before them. Their demand: Bayesian justice for all.

That rallying call derives from the work of Thomas Bayes, an 18th-century British mathematician who showed how to calculate conditional probability – the chance of something being true if its truth depends on other things being true, too. That is precisely the kind of problem that criminal trials deal with as they sift through evidence to establish a defendant's innocence or guilt (see "Bayes on trial", below).

Mathematics might seem a logical fit for the courts, then. Judges and juries, though, all too often rely on gut feeling. A startling example was the rape trial in 1996 of a British man called Dennis John Adams. Adams hadn't been identified in a line-up and his girlfriend had provided an alibi. But his DNA was a 1 in 200 million match to semen from the crime scene – evidence seemingly so damning that any jury would be likely to convict him.

But what did that figure actually mean? Not, as courts and the press often assume, that there was only a 1 in 200 million chance that the semen belonged to someone other than Adams, making his innocence implausible. ›

Bayes on trial

Suppose you have a piece of evidence, E, from a crime scene – a bloodstain, or perhaps a clothing thread – that matches to a suspect. How should it affect your perception or hypothesis, H, of the suspect's innocence?

Bayes's theorem tells you how to work out the probability of H given E. It is: (the probability of H) multiplied by (the probability of E given H) divided by (the probability of E). Or in standard mathematical notation:

$$P(H|E) = P(H) \times P(E|H) / P(E)$$

Say you are a juror at an assault trial, and so far you are 60 per cent convinced the defendant is innocent: P(H) = 0.6. Then you're told that the blood of the defendant and blood found at the crime scene are both type B, which is found in about 10 per cent of people. How should this change your view?

What the forensics expert has given you is the probability that the evidence matches anyone in the general population, given that they are innocent: P(E|H) = 0.1. To apply Bayes's formula and find P(H|E) – your new estimation of the defendant's innocence – you now need the quantity P(E) , the probability that his blood matches that at the crime scene.

This probability actually depends on the defendant's innocence or guilt. If he is innocent, it is 0.1 as it is for anyone else. If he is guilty, however, it is 1, as his blood is certain to match. This insight allows us to calculate P(E) by summing the probabilities of a blood match in the case of innocence (H) or guilt (not H):

$$P(E) = [P(E|H) \times P(H)] + [P(E|\text{not } H) \times P(\text{not } H)]$$
$$= (0.1 \times 0.6) + (1 \times 0.4) = 0.46$$

So according to Bayes's formula, the revised probability of his innocence is:

$$P(H|E) = (0.6 \times 0.1) \div 0.46 = 0.13$$

As you might expect, by this measure the defendant is between four and five times guiltier than you first thought – probably.

CJ GUNTHER/EPA/CORBIS

Allowing juries to rely on "common sense" alone can land innocent people in jail

It actually means there is a 1 in 200 million chance that the DNA of any random member of the public will match that found at the crime scene (see "The prosecutor's fallacy", right). The difference is subtle, but significant. In a population, say, of 10,000 men who could have committed the crime, there would be a 10,000 in 200 million, or 1 in 20,000, chance that someone else is a match too. That still doesn't look good for Adams, but it's not nearly as damning.

So worried was Adams's defence team that the jury might misinterpret the odds that they called in Peter Donnelly, a statistical scientist at the University of Oxford. "We designed a questionnaire to help them combine all the evidence using Bayesian reasoning," says Donnelly.

They failed, though, to convince the jury of the value of the Bayesian approach, and Adams was convicted. He appealed twice unsuccessfully, with an appeal judge eventually ruling that the jury's job was "to evaluate evidence not by means of a formula… but by the joint application of their individual common sense."

But what if common sense runs counter to justice? For David Lucy, a mathematician at Lancaster University in the UK, the Adams judgment indicates a cultural tradition that needs changing. "In some cases, statistical analysis is the only way to evaluate evidence, because intuition can lead to outcomes based upon fallacies," he says.

In 2009 Norman Fenton, a computer scientist at Queen Mary, University of London, who has worked for defence teams in criminal trials, came up with a possible solution. With his colleague Martin Neil, he developed a system of step-by-step pictures and decision trees to help jurors grasp Bayesian reasoning. Once a jury has been convinced that the method works, the duo argue, experts should be allowed to apply Bayes's theorem to the facts of the case as a kind of "black box" that calculates how the probability of innocence or guilt changes as each piece of evidence is presented. "You wouldn't question the steps of an electronic calculator, so why here?" Fenton asks.

It was a controversial suggestion, and it hasn't caught on. Taken to its logical conclusion, it might see the outcome of a trial balance on a single calculation. Working out Bayesian probabilities with DNA and blood matches is all very well, but quantifying incriminating factors such as appearance and behaviour is more difficult. "Different jurors will interpret different bits of evidence differently. It's not the job of a mathematician to do it for them," says Donnelly.

He thinks forensics experts should be schooled in statistics so they can catch errors before they occur. Since cases such as Adams's, that has already begun to happen in the US and UK. Lawyers and jurors, however, still have far less – if any – statistical training.

As the real-life fallacies that follow show, there's no room for complacency. It is not about mathematicians trying to force their way of thinking on the world, says Donnelly: "Justice depends on getting everyone to reason properly with uncertainties." ■

FIVE FALLACIES TO FORGO

It pays to be careful when using statistics as evidence, as these examples from the legal casebook show

1. PROSECUTOR'S FALLACY

"The prosecutor's fallacy is such an easy mistake to make," says Ian Evett of Principal Forensic Services, a UK forensics company. It confuses two subtly different probabilities that Bayes's formula distinguishes: P(H|E), the probability that someone is innocent if they are a match to a piece of evidence, and P(E|H), the probability that someone is a match to a piece of evidence if they are innocent (see "Bayes on trial", previous page). The first probability is what we would like to know; the second is what forensics usually tells us.

Unfortunately, even professionals sometimes mix them up. In the 1991 rape trial of Andrew Deen in Manchester, UK, for example, an expert witness agreed on the basis of a DNA sample that "the likelihood of [the source of the semen] being any other man but Andrew Deen [is] 1 in 3 million."

That was wrong. One in 3 million was the likelihood that any innocent person in the general population had a DNA profile matching that extracted from semen at the crime scene – in other words, P(E|H). With around 60 million people in the UK, a fair few people will share that profile.

Depending on how many of them might plausibly have committed the crime, the probability of Deen being innocent even though he was a match, or P(H|E), was actually far greater that 1 in 3 million. Deen's conviction was quashed on appeal, leading to a flurry of similar appeals that have had varying success.

2. ULTIMATE ISSUE ERROR

The prosecution in the Deen case stopped just short of compounding their probabilistic fallacy. In the minds of the jury, though, it probably morphed into the "ultimate issue" error: explicitly equating the (small) number P(E|H) with a suspect's likelihood of innocence.

In Los Angeles in 1968, the ultimate issue error sent Malcolm Collins and his wife Janet to jail. At first glance, the circumstances of the case left little room for doubt: an elderly lady had been robbed by a white woman with blonde hair and a black man with a moustache, who had both fled in a yellow car. The chances of finding a similar interracial couple matching that description were 1 in 12 million, an expert calculated.

The police were convinced, and without much deliberation so was the jury. They assumed that there was a 1 in 12 million chance that the couple were not the match, and that this was also the likelihood of their innocence.

They were wrong on both counts. In a city such as Los Angeles, with millions of people of all races living in it or passing through, there could well be at least one other such couple, giving the Collinses an evens or better chance of being innocent. Not to mention that the description itself may have been inaccurate – facts that helped reverse the guilty verdict on appeal.

3. BASE-RATE NEGLECT

Anyone looking to DNA profiling for a quick route to a conviction should recognise that genetic evidence can be shaky. Even if the odds of finding another genetic match are 1 in a billion, in a world of 7 billion, that's another seven people with the same profile.

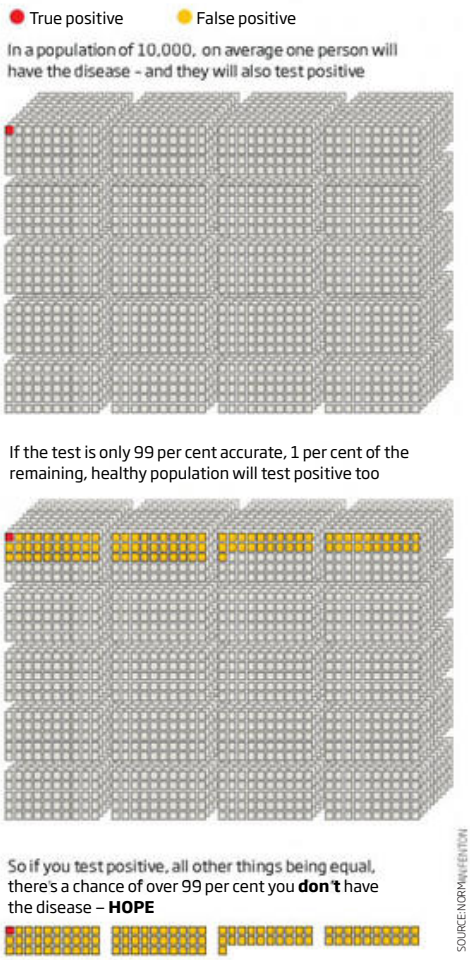
Fortunately, circumstantial and forensic evidence often quickly whittle down the pool of suspects. But neglecting your "base rate" – the pool of possible matches – can have you leap to false conclusions, not just in the courtroom.

Picture yourself, for example, in the doctor's surgery. You have just tested positive for a terminal disease that afflicts 1 in 10,000. The test has an accuracy of 99 per cent. What's the probability that you actually have the disease?

It is in fact less than 1 per cent. The reason is the sheer rarity of the disease, which means that even with a 99 per cent accurate test, false positives will far outweigh real ones (see diagram, above). That's why it is so important to carry out further tests to narrow down the odds. We lay people are not the only ones stumped by such counter-intuitive results: surveys show that 85 to 90 per cent of health professionals get it wrong too.

Don't panic

You've just been diagnosed with a rare condition that afflicts 1 in 10,000. The test is 99 per cent certain. Hope or despair?



SOURCE: NORMAN FENTON

4. DEFENDANT'S FALLACY

It's not just prosecutors who can fiddle courtroom statistics to their advantage: defence lawyers have also been known to cherry-pick probabilities.

In 1995, for example, former American football star O.J. Simpson stood trial for the murder of his ex-wife, Nicole Brown, and her friend. Years before, Simpson had pleaded no contest to a charge of domestic violence against Brown. In an attempt to downplay that, a consultant to Simpson's defence team, Alan Dershowitz, stated that fewer than 1 in 1000 women who are abused by their husbands or boyfriends end up murdered by them.

That might well be true, but it was not the most relevant fact, as John Allen Paulos, a mathematician at Temple University in Philadelphia, Pennsylvania, later showed. As a Bayesian calculation taking in all the pertinent facts reveals, it is trumped by the 80 per cent likelihood that, if a woman is abused and later murdered, the culprit was her partner.

That may not be the whole story either, says criminologist William Thompson of the University of California, Irvine. If more than 80 per cent of all murdered women, abused or not, are killed by their partner, "the presence of abuse may have no diagnostic value at all".

5. DEPENDENT EVIDENCE FALLACY

Sometimes, mathematical logic flies out of the courtroom window long before Bayes can even be applied – because the probabilities used are wrong.

Take the dependent evidence fallacy, which was central to a notorious miscarriage of justice in the UK. In November 1999, Sally Clark was convicted of smothering her two children as they slept. Paediatrician Roy Meadow testified that the odds of both dying naturally by sudden infant death syndrome (SIDS), or cot death, were 1 in 73 million. He arrived at this figure by multiplying the individual probability of SIDS in a family such as Clark's – 1 in 8500 – by itself, as if the two deaths were independent events.

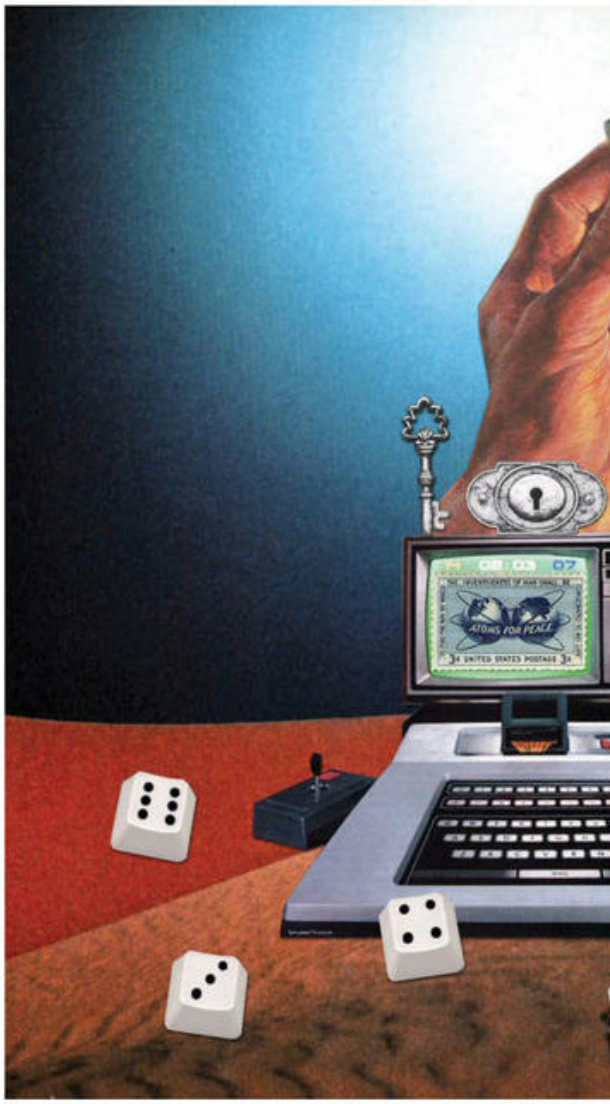
But why should they be? "There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely," the Royal Statistical Society explained during an appeal.

"Even three eminent judges didn't pick up on the mistake," says Ray Hill of the University of Salford, who worked for the defence team. He estimated that if one sibling dies of SIDS, the chance of another dying is as high as 1 in 60. Bayesian reasoning then produces a probability of a double cot death of around 1 in 130,000. With hundreds of thousands of children born each year in the UK, there's bound to be a double cot death from time to time.

Clark was eventually freed on appeal in 2003. Her case had a lasting effect, leading to the review of many similar cases. "I'm not aware of any cases of multiple cot deaths reaching the courts in recent years," says Hill. Clark herself never recovered from her ordeal, however. She was found dead at her home in 2007, ultimately a victim of statistical ignorance.

THINK OF A NUMBER

Chances are it won't be a random number, says **Michael Brooks**



MADS HAAHR is in no doubt. "Generating randomness is not a task that should be left to humans," he says.

You might expect him to say that. A computer scientist at Trinity College Dublin, he is the creator of a popular online random number generator, hosted at random.org. But he has a point.

Human brains are wired to spot and generate patterns. That is useful when it's all about seeing predators on the savannah before they see you, but it handicaps us when we need to think in random and unpredictable ways (see "The mathematician", right). That's a problem, because true randomness is a useful thing to have. Random numbers are used in cryptography, computing, design and many other applications. Our inability to "do" random means that we usually have to outsource it to machines.

But relying on outside sources of randomness has its own problems. The first dice for divination and gaming were six-sided bones from the heels of sheep, with numbers carved into the faces. The shape made some numbers more likely to appear than others, giving a decisive advantage to those who understood its properties.

Suspicion about the reliability of randomness generators remains with modern equivalents like casino dice, roulette wheels or lottery balls. But it is online where it really matters. Generating random strings of numbers is essential not just for gambling games or shuffling songs on your iPod, but also to produce unguessable keys used to encrypt sensitive digital information. "I don't think people are very conscious of how important randomness is for the security of their data," says Haahr.

And it takes more than programming. You can't just give computers rules to create random numbers; that wouldn't be random. Instead you might use an algorithm to "seed" a random-looking output from a smaller, unpredictable input: use the date and time to determine which random digits to extract from a random number string such as π, say, and work from there. The problem is that such "pseudorandom" numbers are limited by the input, and tend to repeat non-randomly after a certain time in a way that is guessable if you see enough of them.

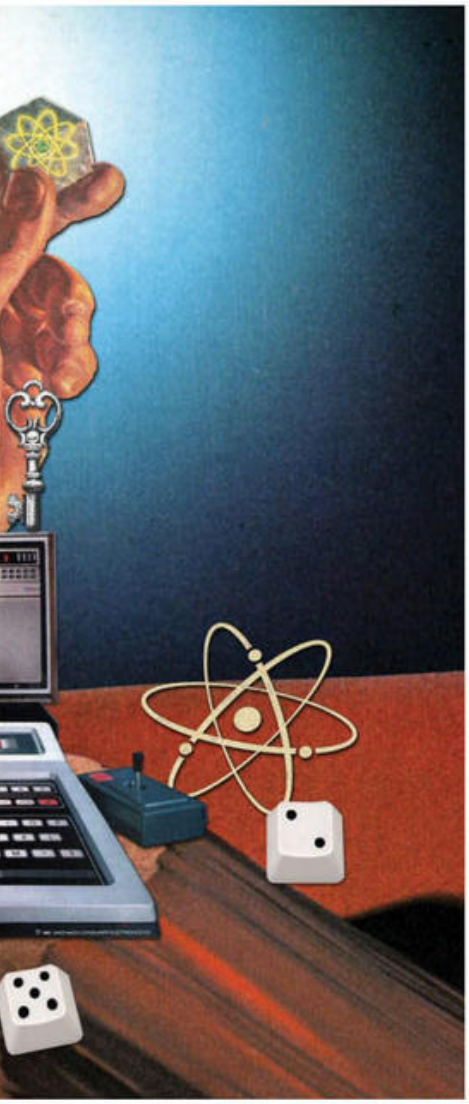
An alternative is to hook up your computer to some source of "true" randomness. In the 1950s, the UK Post Office wanted to generate industrial quantities of random numbers to pick the winners of its Premium Bonds lottery. The job fell to the designers of the Colossus computer, developed to crack Nazi Germany's Enigma codes. They created ERNIE, the Electronic Random Number Indicator Equipment, which harnessed the chaotic trajectories of electrons passing through neon tubes to produce a randomly timed series of electronic pulses.

"OUR BRAINS HANDICAP US WHEN WE NEED TO THINK IN RANDOM AND UNPREDICTABLE WAYS"

ERNIE is now in his fourth iteration and is a simpler soul, relying on thermal noise from transistors to generate randomness. Many modern computing applications use a similar source, collected using on-chip generating units such as Intel's RdRand and Via's Padlock. Haahr's generator takes its seed from intrinsically noisy atmospheric processes.

Two problems remain. First, with enough computing power anyone can, in theory, reconstruct the processes of classical physics that created the random numbers. Second, and more practically, random number generators based solely on physical processes often can't produce random bits fast enough.

Many systems, such as the Unix-based platforms used by Apple, get round the first problem by combining the output from on-chip randomness generators with the contents of an "entropy pool", filled with other random contributions. This could be anything from thermal



EUGENIA LOLI

noise in devices connected to the computer to the random timings of the user's keyboard strokes. The components are then combined using a "hash function" to generate a single random number. Hash functions are the mathematical equivalent of stirring ink into water: there's no known way to work out what the set of inputs was, given the number the function spits out.

That doesn't mean there couldn't be in the future – and there's still the speed problem. The workaround is generally to use a physical random number generator only as a seed for a program that generates a more abundant flow.

Then we are back with the algorithm problem. The precise nature of the methods these programs use is proprietary, but in 2013, security analysts raised concerns that the US National Security Agency knew the internal workings of one such generator, called Dual_EC_DRBG, potentially allowing them to break encryptions that relied on it. If you're just playing online games, that's not a big problem. But when making multibillion-dollar financial transactions, or encrypting sensitive documents, a suspicion that you are being watched is a bigger deal.

Gaming the system

Such difficulties lead some researchers to suggest we will never have an uncrackable source of randomness as long as we rely on the classical world, where randomness is not intrinsic, but down to who has what information (see "Random reality", page 106). For safer encryption, we must turn to quantum physics, where things truly do seem random. Instead of a coin toss, you might ask whether a photon hitting a half-silvered mirror passed through it or was reflected. Instead of rolling a die, you might present an electron with a choice of six circuits to pass through. "As a mathematician, I like my randomness to come with proof, and quantum random numbers give us that," says Carl Miller of the University of Michigan in Ann Arbor. "It's unique in that respect."

Cryptographic systems that exploit the vagaries of quantum theory for more secure communication do exist. But they are not the last word in security. Extracting quantum randomness always involves someone making non-random choices about equipment, measurements and such like. The less-than-perfect efficiency of photon detectors used in some methods could also provide a back door through which non-randomness can slip in.

One way out that is still under investigation might be to amplify quantum randomness so you always have more of it than anyone can hack. Ways exist in theory to turn n random bits into 2^n bits of pure randomness, and also to launder bits to remove any correlation with the device that first made them.

Such device-independent quantum random number generation is just the latest development in our search for true randomness. Chances are, this too will soon become reality – only then for someone to find a way to game it. With humans forever in the mix, it could be that we'll always be searching for randomness we can rely on. ■

THE MATHEMATICIAN
DAVID HAND

Emeritus professor, Imperial College London

In your recent book *The Improbability Principle*, you state that extremely unlikely events are commonplace. How so?

At first glance, it sounds like a contradiction: if something is highly improbable, how can it possibly be commonplace? But as you dig deeper you see it is not a contradiction, and that you should expect what appear to be extremely improbable events to occur quite often. The principle itself is really an interweaving of five fundamental laws.

Could you give an example of one of those laws?

Take the law of truly large numbers. The most obvious example of this is a lottery. In a 49-ball game you have a 1 in 14 million chance of winning if you buy just one ticket. But of course if you get enough people buying enough tickets it becomes almost inevitable that somebody somewhere will win. Another example is the chance of being struck by lightning. Around the world there's a 1 in 300,000 chance of being killed by lightning in any one year. The rational thing is to behave as if it's not going to happen to you. But there are 7 billion people in the world, so there are a lot of opportunities for it to happen. In fact the chance that no one will be killed is about $10^{-10,133}$. So we should expect to see someone killed. In fact about 24,000 people every year are killed by lightning, and about 10 times that many are injured.

People often notice coincidences and patterns that aren't really there. Why?

Our ancestors survived in the world because they identified patterns: if you responded to movements in the grass you could avoid being killed by an approaching tiger. So there's an evolutionary reason. But a lot of what look like patterns in data just appear by chance.

Definitely not maybe

When it comes to explaining the world, probability is as much use as flat-Earth theory, asserts physicist **David Deutsch**

Probability theory was devised by gamblers hoping to win more money

PROBABILITY theory is a quaint little piece of mathematics. It is about sets of non-negative numbers that are attached to actual and possible physical events, that sum to 1 and that obey certain rules. It has numerous practical applications.

So does flat-Earth theory: for instance, it's an excellent approximation when laying out your garden.

Science abandoned the misconception that Earth extends over an infinite plane, or has edges, millennia ago. Probability insinuated itself into physics relatively recently, yet the idea that the world actually follows probabilistic rules is even more misleading than saying Earth is flat. Terms such as "likely", "probable", "typical" and "random", and statements assigning probabilities to physical events are incapable of saying anything about what actually will happen.

We are so familiar with probability statements that we rarely wonder what "*x* has a probability of ½" actually asserts about the world. Most physicists think that it means something like: "If the experiment is repeated infinitely often, half of the time the outcome will be *x*." Yet no one repeats an experiment infinitely often. And from that statement about an infinite number of outcomes, nothing follows about any finite number of outcomes. You cannot even define probability statements as being about what will happen in the long run. They only say what will *probably* happen in the long run.

The awful secret at the heart of probability theory is that physical events either happen or they don't: there's no such thing in nature as probably happening. Probability statements aren't factual assertions at all.

The theory of probability as a whole is irretrievably "normative": it says what ought to happen in certain circumstances and then presents us with a set of instructions. It is normative because it commands that very high probabilities, such as "the probability of *x* is near 1", should be treated almost as if they were "*x* will happen". But such a normative rule has no place in a scientific theory, especially not in physics. "There was a 99 per cent chance of sunny weather yesterday" does not mean "It was sunny".

It all began quite innocently. Probability and associated ideas such as randomness didn't originally have any deep scientific purpose. They were invented in the 16th and 17th centuries by people who wanted to win money at games of chance.

Gaming the system

To discover the best strategies for playing such games, they modelled them mathematically. True games of chance are driven by chancy physical processes such as throwing dice or shuffling cards. These have to be unpredictable (having no known pattern) yet equitable (not favouring any player over another). The three-card trick, for example, does not qualify: the conjurer deals the cards unpredictably (to the onlooker) but not equitably. A roulette wheel that indicates each of its numbers in turn, meanwhile, behaves equitably but predictably, so equally cannot be used to play a real game of roulette.

Earth was known to be spherical long before physics could explain how that was possible. Similarly, before game theory, mathematics could not yet accommodate an unpredictable, equitable sequence of numbers, so game theorists had to invent mathematical randomness and probability. They analysed games as if the chancy elements were generated by "randomisers": abstract devices generating random sequences, with uniform probability. Such sequences are indeed unpredictable and equitable – but also have other, quite counter-intuitive properties.

For a start, no finite sequence can be truly random. To expect fairly tossed dice to be less likely to come up with a double after a long sequence of doubles is a falsehood known as the gambler's fallacy. But if you know that a finite sequence is equitable – it has an equal number of 1s and 0s, say – then towards the end, knowing what came before does make it easier to predict what must come next.

A second objection is that because classical physics is deterministic, no classical mechanism can generate a truly random sequence. So why did game theory work? Why

LUCY RIDGES/MILLENNIUM IMAGES, UK

was it able to distinguish useful maxims, such as "never draw to an inside straight" in poker, from dangerous ones such as the gambler's fallacy? And why, later, did it enable true predictions in countless applications, such as Brownian motion, statistical mechanics and evolutionary theory? We would be surprised if the four of spades appeared in the laws of physics. Yet probability, which has the same provenance as the four of spades but is nonsensical physically, seems to have done just that.

The key is that in all of these applications, randomness is a very large sledgehammer used to crack the egg of modelling fair dice, or Brownian jiggling with no particular pattern, or mutations with no intentional design. The conditions that are required to model these situations are awkward to express mathematically, whereas the condition of randomness is easy, given probability theory. It is unphysical and far too strong, but no matter. One can argue that replacing the dice with a mathematical randomiser would not change the strategy of an ideally rational dice player – but only if the player assumes that pesky normative rule that a very high probability of something happening should be treated as a statement that it will happen.

So the early game theorists never did quite succeed at finding ways of winning at games of chance: they only found ways of probably winning. They connected those with reality by supposing the normative rule that "very probably winning" almost equates to "winning". But every gambler knows that probably winning alone will not pay the rent. Physically, it can be very unlike actually winning. We must therefore ask what it is about the physical world that nevertheless makes obeying that normative rule rational.

You may have wondered when I mentioned the determinism of classical physics whether quantum theory solves the problem. It does, but not in the way one might expect. Because quantum physics is deterministic too.

"Probability and randomness are large sledgehammers to crack some small eggs"

Indeterminism – what Einstein called "God playing dice" – is an absurdity introduced to deny the implication that quantum theory describes many parallel universes. But it turns out that under deterministic, multi-universe quantum theory, the normative rule follows from ordinary, non-probabilistic normative assumptions such as "if x is preferable to y, and y to z, then x is preferable to z".

You could conceive of Earth as being literally flat, as people once did, and that falsehood might never adversely affect you. But it would also be quite capable of destroying our entire species, because it is incompatible with developing technology to avert, say, asteroid strikes. Similarly, conceiving of the world as being literally probabilistic may not prevent you from developing quantum technology. But because the world isn't probabilistic, it could well prevent you from developing a successor to quantum theory. In particular, constructor theory – a framework that I have advocated for fundamental physics, within which I expect successors to quantum theory to be developed – is deeply incompatible with physical randomness.

It is easy to accept that probability is part of the world, just as it's easy to imagine Earth as flat when in your garden. But this is no guide to what the world is really like, and what the laws of nature actually are. ■

ANDREW HEM

THE HARDEST PROBLEM

Solving it would net someone a $1 million prize, yet to the rest of us it would be priceless. **Jacob Aron** reports

"DEAR Fellow Researchers, I am pleased to announce a proof that P is not equal to NP, which is attached in 10pt and 12pt fonts." So began an email sent in August 2010 to a group of leading computer scientists by Vinay Deolalikar, a mathematician at Hewlett-Packard Labs in Palo Alto, California.

It was an incendiary claim. Deolalikar was saying he had cracked the biggest problem in computer science, a question concerning the complexity and fundamental limits of computation. Answer the question "Is P equal to NP?" with a yes, and you could be looking at a world transformed, where planes and trains run on time, accurate electronic translations are routine, the molecular mysteries of life are revealed at the click of a mouse – and secure online shopping becomes fundamentally impossible. Answer it with a no, as Deolalikar claimed to have done, and it would suggest that some problems are too irreducibly complex for computers to solve exactly.

One way or another, then, we would like an answer. But it has been frustratingly slow in coming. "It's turned out to be an incredibly hard problem," says Stephen Cook of the University of Toronto, Canada, the computer scientist who first formulated it in May 1971.

The importance of P = NP? was emphasised in 2000, when the privately funded Clay Mathematics Institute in Cambridge, Massachusetts, named it as one of seven "Millennium Prize" problems, with a $1 million bounty for whoever solves it. Since then, Cook and other researchers in the area have regularly received emails with purported solutions. Gerhard Woeginger of the Eindhoven University of Technology in the Netherlands even maintains an online list of them. "We've seen hundreds of attempts, and we can more or less classify them according to the mistakes the proof makes," he says.

Deolalikar's attempt seemed different. For a start, it came from an established researcher rather than one of the legion of amateurs hoping for a pop at glory and a million dollars. Also, Cook initially gave it a cautious thumbs up, writing that it seemed to be a "relatively serious claim". That led it to spread across the web and garnered it widespread attention in the press.

In the end, though, it was a false dawn. "It probably didn't deserve all the publicity," says Neil Immerman, a computer scientist at the University of Massachusetts, Amherst. He was one of an army of researchers who, working in an informal online collaboration, soon exposed fundamental flaws in the proof.

The consensus now is that his proof – like all attempts before it – is unfixable. And so the mystique of P = NP? remains unbroken almost half a century after it was first formulated. But what is the problem about? Why is it so important, and what happens if it is answered one way or the other? ›

WHAT IS P?

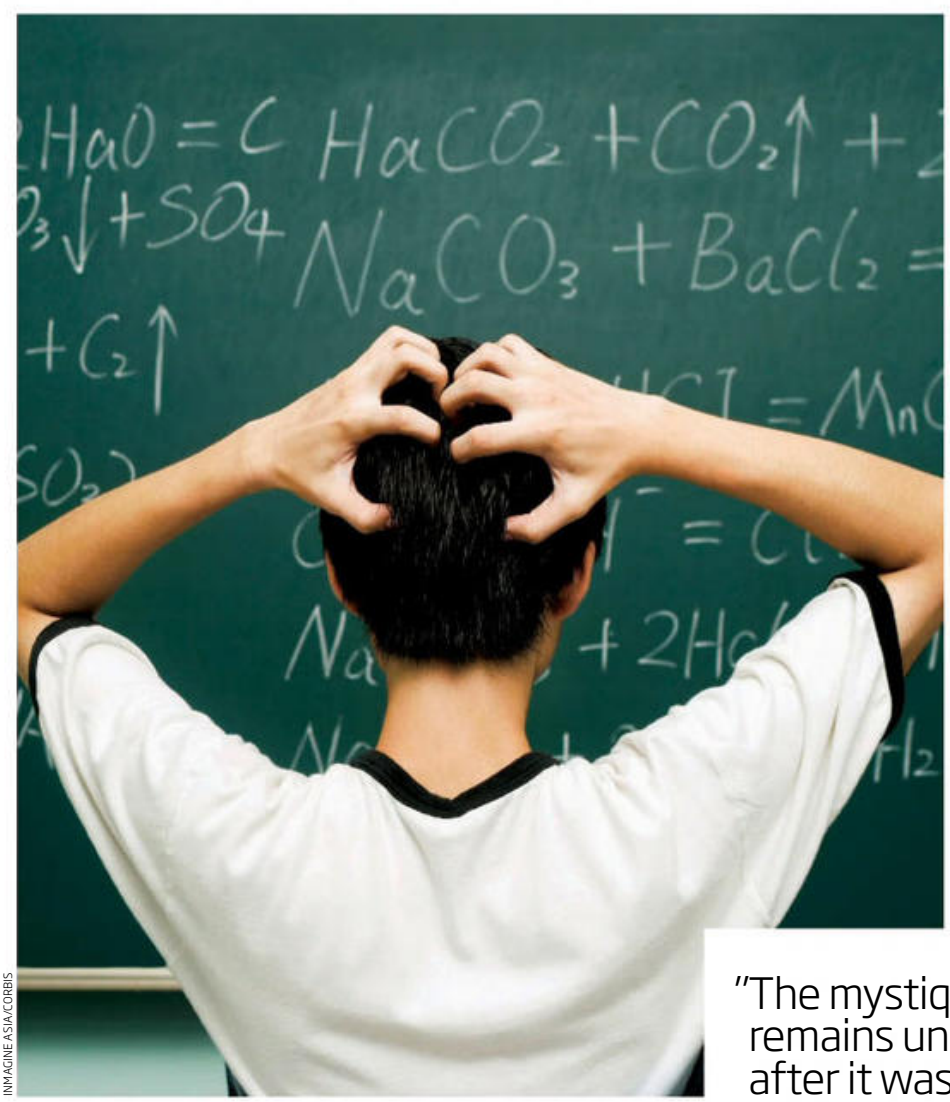
Computing power, we are accustomed to think, is the gift that keeps on giving. Every year or two that goes by sees roughly a doubling in our number-crunching capabilities – a march so relentless that it has acquired the label "Moore's law", after the Intel researcher, Gordon Moore, who first noted the trend back in the 1960s.

Talk of the ultimate limits of computation tends to be about how many transistors, the building blocks of microprocessors, we can cram onto a conventional silicon chip – or whatever technology or material might replace it. P = NP? raises the spectre that there is a more fundamental limitation, one that lies not in hardware but in the mechanics of computation itself.

P = NP should not be mistaken for a simple algebraic equation – if it were, we could just have N = 1 and claim the Clay institute's million bucks. P and NP are examples of "complexity classes", categories into which problems can be slotted depending on how hard it is to unravel them using a computer.

Solving any problem computationally depends on finding an algorithm, the step-by-step set of mathematical instructions that leads us to the answer. But how much number-crunching does an algorithm need? That depends on the problem.

The P class of problems is essentially the easy ones: an algorithm exists to solve them in a "reasonable" amount of time. Imagine looking to see if a particular number appears in a list. The simplest solution is a "linear search" algorithm: you check each number in turn until you find the right one. If the list has n numbers – the "size" of the problem – this algorithm takes at most n steps to search it, so its complexity is proportional to n. That counts as reasonable. So too do things that take a little more computational muscle – for instance, the manual multiplication of two n-digit numbers, which takes about n^2 steps. A pocket calculator will still master that with ease, at least for relatively small values of n. Any problem of size n whose solution requires n to the power of something (n^x) steps is relatively quick to crack. It is said to be solvable in "polynomial time", and is denoted P.



Finding structural similarities of chemical compounds is an NP-complete problem

IN IMAGINE ASIA/CORBIS

WHAT IS NP?

Not all problems are as benign. In some cases, as the size of the problem grows, computing time increases not polynomially, as n^x, but exponentially, as x^n – a much steeper increase. Imagine, for example, an algorithm to list out all possible ways to arrange the numbers from 1 to n. It is not difficult to envisage what the solutions are, but even so the time required to list them rapidly runs away from us as n increases. Even proving a problem belongs to this non-polynomial class can be difficult, because you have to show that

"The mystique of the problem remains unbroken half a century after it was first formulated"

absolutely no polynomial-time algorithm exists to solve it.

That does not mean we have to give up on hard problems. With some problems that are difficult to solve in a reasonable time, inspired guesswork might still lead you to an answer whose correctness is easy to verify. Think of a sudoku puzzle. Working out a solution can be fiendishly difficult, even for a computer, but presented with a completed puzzle, it is easy to check that it fulfils the criteria for being a valid answer (see diagram, below). Problems whose solutions are hard to come by but can be verified in polynomial time make up the complexity class called NP, which stands for non-deterministic polynomial time.

Constructing a valid sudoku grid – in essence asking the question "Can this number fill this space?" over and over again for each space and each number from 1 to 9, until all spaces are compatibly filled – is an example of a classic NP problem, the Boolean satisfiability problem. Processes such as using formal logic to check software for errors and deciphering the action of gene regulatory networks boil down to similar basic satisfiability problems.

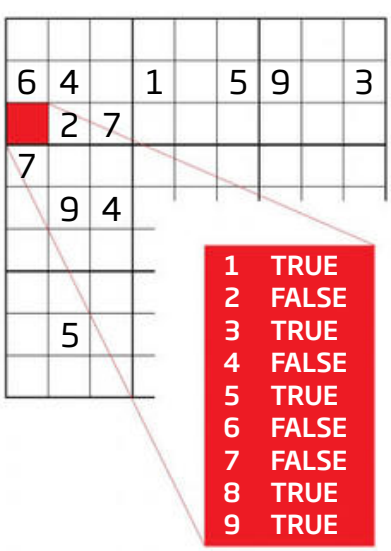
And here is the nub of the P = NP? problem. All problems in the set P are also in the set NP: if you can easily find a solution, you can easily verify it. But is the reverse true? If you can easily verify the solution to a problem, can you also easily solve it – is every problem in the set NP also in P? If the answer to that question is yes, the two sets are identical: P is equal to NP, and the world of computation is irrevocably changed – sudoku becomes a breeze for a machine to solve, for starters. But before we consider that possibility, what happens if P is found to be not equal to NP?

WHAT IF P ≠ NP?

In 2002, William Gasarch, a computer scientist at the University of Maryland in College Park, asked 100 of his peers what ›

Can't get no satisfaction

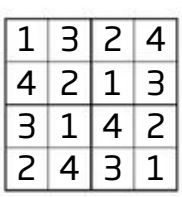
Constructing a valid Sudoku grid is an example of a computational problem known as a Boolean satisfiability problem



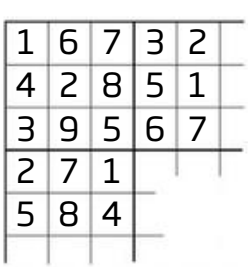
Given an incomplete sudoku grid, finding a viable solution amounts to evaluating a boolean "true" or "false" answer for each empty square – whether each number from 1 to 9 can fit there – and solving it iteratively until no ambiguities remain

Satisfiability problems are "NP-hard": as the size of the problem increases, it takes far more computational muscle to find a solution than to check it

For a 1x1 grid the (only possible) solution is trivial



For a 4x4 grid, generating a viable solution still takes little computational effort



A 9x9 grid takes considerably more effort to construct – but it is still relatively easy to check the solution is right

they thought the answer to the P = NP? question would be. "You hear people say offhand what they think of P versus NP. I wanted to get it down on record," he says. P ≠ NP was the overwhelming winner, with 61 votes. Only nine people supported P = NP, some, they said, just to be contrary. The rest either had no opinion or deemed the problem impossible to solve.

If the majority turns out to be correct, and P is not equal to NP – as Deolalikar suggested – it indicates that some problems are by their nature so involved that we will never be able to crunch through them. If so, the proof is unlikely to make a noticeable difference to you or me. In the absence of a definitive answer to P = NP?, most computer scientists already assume that some hard problems cannot be solved exactly. They concentrate on designing algorithms to find approximate solutions that will suffice for most practical purposes. "We'll be doing exactly the same as we're currently doing," says Woeginger.

Proving P ≠ NP would still have practical consequences. For a start, says Immerman, it would shed light on the performance of the latest computing hardware, which splits computations across multiple processors operating in parallel. With twice as many processors, things should run twice as fast – but for certain types of problem they do not. That implies some kind of limitation to computation, the origin of which is unclear. "Some things seem like they're inherently sequential," says Immerman. "Once we know that P and NP are not equal, we'll know why."

It could also have an impact on the world of cryptography. Most modern encryption relies on the assumption that breaking a number down to its prime-number factors is hard. This certainly looks like a classic NP problem: finding the prime factors of 304,679 is hard, but it's easy enough to verify that they are 547 and 557 by multiplying them together. Real encryption uses numbers with hundreds of digits, but a polynomial-time algorithm for solving NP problems would crack even the toughest codes.

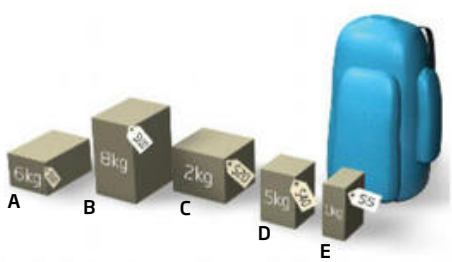
So would P ≠ NP mean cryptography is secure? Not quite. In 1995, Russell Impagliazzo of the University of California, San Diego, sketched out five possible outcomes of the P = NP? debate. Four of them were shades-of-grey variations on a world where P ≠ NP. For example, we might find some problems where even though the increase in complexity as the problem grew is technically exponential, there are still relatively efficient ways of finding solutions. If the prime-number-factoring problem belongs to this group, cryptography's security could be vulnerable, depending where reality lies on Impagliazzo's scale. "Proving P is not equal to NP isn't the end of our field, it's the beginning," says Impagliazzo.

Ultimately, though, it is the fifth of Impagliazzo's worlds that excites researchers the most, however unlikely they deem it. It is "Algorithmica" – the computing nirvana where P is indeed equal to NP.

> "Proving P = NP would have the curious side effect of rendering mathematicians redundant"

Pack it in

In the knapsack problem, your bag can only hold a certain weight, here 20 kg. Which objects should you pack to maximise the value of the contents?



An obvious strategy is to pack the objects with the highest value per unit-weight until you reach the weight limit

	Value	Kilograms	Value/kg
A	$90	6	$15/kg
B	$100	8	$12.50/kg
C	$20	2	$10/kg
D	$40	5	$8/kg
E	$5	1	$5/kg

Pack A, then B, then C. D would be too heavy, so leave it and add E

TOTAL $215

But pack A, B, D then E and you can carry goods worth $20 more

TOTAL $235

For large numbers of objects, there is no easy way to arrive at this optimal solution – unless it can be proved that P = NP

WHAT IF P = NP?

If P = NP, the revolution is upon us. "It would be a crazy world," says Immerman. "It would totally change our lives," says Woeginger.

That is because of the existence, proved by Cook in his seminal 1971 paper, of a subset of NP problems known as NP-complete. They are the executive key to the NP washroom: find an algorithm to solve an NP-complete problem, and it can be used to solve any NP problem in polynomial time.

Lots of real-world problems are known to be NP-complete. Satisfiability problems are one example; versions of the knapsack problem (left), which deals with optimal resource allocation, and the notorious travelling salesman problem (above right) are others. This problem aims to find the shortest-distance route for visiting a series of points and returning to the starting point, an issue of critical interest in logistics and elsewhere.

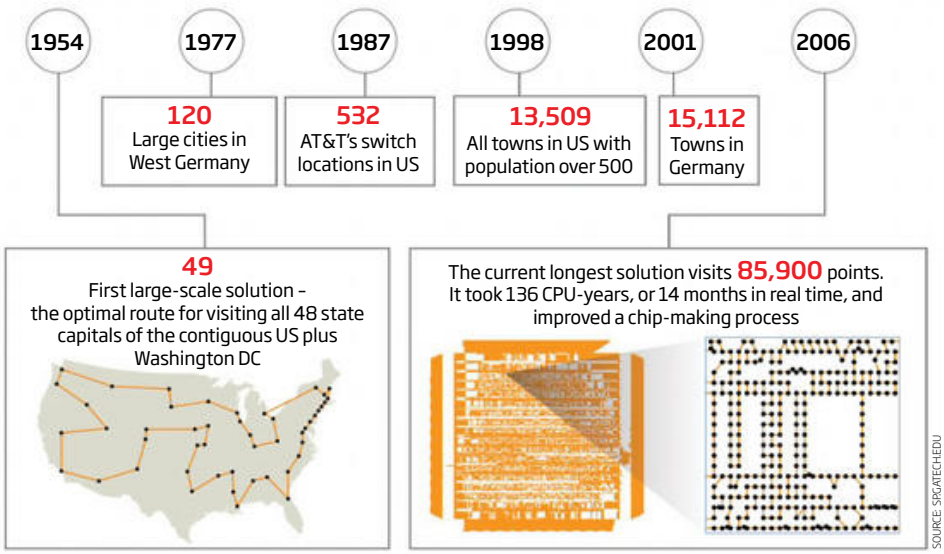
If we could find a polynomial-time algorithm for any NP-complete problem, it would prove that P = NP, since all NP problems would then be easily solvable. The existence of such a universal computable solution would allow the perfect scheduling of transport, the most efficient distribution of goods, and manufacturing with minimal waste – a leap and a bound beyond the "seems-to-work" solutions we now employ.

It could also lead to algorithms that perform near-perfect speech recognition and language translation, and that let computers process visual information as well as any human can. "Basically, you would be able to compute anything you wanted," says Lance Fortnow, a computer scientist at Northwestern University in Evanston, Illinois.

A further curious side effect would be that of rendering mathematicians redundant. "Mathematics would be largely

The route master

Finding the shortest route that takes in multiple locations – the travelling salesman problem – is an important issue in real-world logistics. In the absence of an efficient general algorithm, finding solutions is a question of throwing computing power at it. As computers have sped up, so the complexity of cases solved has grown



mechanisable," says Cook. Because finding a mathematical proof is difficult, but verifying one is relatively easy, in some way maths itself is an NP problem. If P = NP, we could leave computers to churn out new proofs.

"It could also lead to algorithms for near-perfect speech recognition and translation"

WHAT IF P = NP, BUT THERE IS NO ALGORITHM?

There is an odd wrinkle in visions of a P = NP world: that we might prove the statement to be true, but never be able to take advantage of that proof. Mathematicians sometimes find "non-constructive proofs" in which they show that a mathematical object exists without actually finding it. So what if they could show that an unknown P algorithm exists to solve a problem thought to be NP? "That would technically settle the problem, but not really," says Cook.

There would be a similar agonising limbo if a proof that P = NP is achieved with a universal algorithm that scales in complexity as *n* to the power of a very large number. Being polynomial, this would qualify for the Clay institute's $1 million prize, but in terms of computability it would not amount to a hill of beans.

ARE WE LIKELY TO HAVE ANY DEFINITIVE ANSWER SOON?

When can we expect the suspense to be over, one way or another? Probably not so soon. "Scientometrist" Samuel Arbesman of the Harvard Medical School in Boston, Massachusetts, predicts that a solution of "P = NP?" is not likely to arrive before 2024. In Gasarch's 2002 poll of his peers, only 45 per cent believed it would be resolved by 2050. "I think people are now more pessimistic," he says, "because after 10 years there hasn't been that much progress." He adds that he believes a proof could be up to 500 years away.

Others find that excessively gloomy, but all agree there is a mammoth task ahead. "The current methods we have don't seem to be giving us progress," says Fortnow. "All the simple ideas don't work, and we don't know where to look for new tools," says Woeginger. Part of the problem is that the risk of failure is too great. "If you have already built up a reputation, you don't want to publish something that makes other people laugh," says Woeginger.

Some are undeterred. One is Ketan Mulmuley at the University of Chicago. Fellow researchers say his approach looks promising. In essence it involves translating the P = NP? problem into more tractable problems in algebraic geometry, the branch of mathematics that relates shapes and equations. But it seems even Mulmuley is not necessarily anticipating a quick success. "He expects it to take well over 100 years," says Fortnow.

Ultimately, though, Mulmuley's tactic of connecting P = NP? to another, not obviously related, mathematical area seems the most promising line of attack. It has been used before: in 1995, the mathematician Andrew Wiles used work linking algebraic geometry and number theory to solve another high-profile problem, Fermat's last theorem. "There were a lot of building blocks," says Fortnow. "Then it took one brilliant mind to make that last big leap."

Woeginger agrees: "It will be solved by a mathematician who applies methods from a totally unrelated area, that uses something nobody thinks is connected to the P versus NP question." Perhaps the person who will settle P = NP? is already working away in their own specialised field, just waiting for the moment that connects the dots and solves the world's hardest problem. ■

SIMON GARDINER

The world maker

Is time running out for the clever piece of maths that simplifies the complexities of modern life, asks Richard Elwes

YOU might not have heard of the algorithm that runs the world. Few people have, though it can determine much that goes on in our day-to-day lives: the food we have to eat, our schedule at work, when the train will come to take us there. Somewhere, in some server basement right now, it is probably working on some aspect of your life tomorrow, next week, in a year's time.

Perhaps ignorance of the algorithm's workings is bliss. The door to Plato's Academy in ancient Athens is said to have borne the legend "let no one ignorant of geometry enter". That was easy enough to say back then, when geometry was firmly grounded in the three dimensions of space our brains were built to cope with. But the algorithm operates in altogether higher planes. Four, five, thousands or even many millions of dimensions: these are the unimaginable spaces the algorithm's series of mathematical instructions was devised to probe.

Perhaps, though, we should try a little harder to get our heads round it. Because powerful though it undoubtedly is, the algorithm is running into a spot of bother. Its mathematical underpinnings, despite not yet being structurally unsound, are beginning to crumble at the edges. With so much resting on it, the algorithm may not be quite as dependable as it once seemed.

To understand what all this is about, we must first delve into the deep and surprising ways in which the abstractions of geometry describe the world around us. Ideas about such connections stretch at least as far back as Plato, who picked out five 3D geometric shapes, or polyhedra, whose perfect regularity he thought represented the essence of the cosmos. The tetrahedron, cube, octahedron and 20-sided icosahedron embodied the "elements" of fire, earth, air and water, and the 12-faced dodecahedron the shape of the universe itself.

Things have moved on a little since then. Theories of physics today regularly invoke strangely warped geometries unknown to Plato, or propose the existence of spatial dimensions beyond the immediately obvious three. Mathematicians, too, have reached for ever higher dimensions, extending ideas about polyhedra to mind-blowing "polytopes" with four, five or any number of dimensions.

A case in point is a law of polyhedra proposed in 1957 by the US mathematician Warren Hirsch. It stated that the maximum number of edges you have to traverse to get between two corners on any polyhedron is never greater than the number of its faces minus the number of dimensions in the problem, in this case three. The two opposite corners on a six-sided cube, for example, are separated by exactly three edges, and no pair of corners is four or more apart.

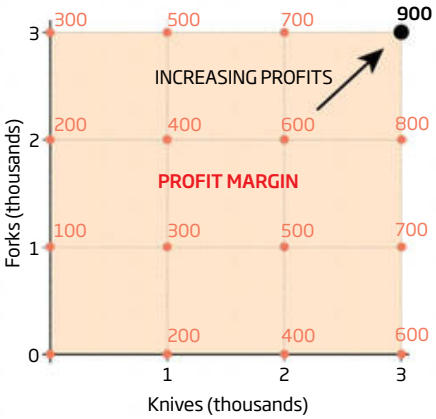
Hirsch's rule holds true for all 3D polyhedra. But it has never been proved generally for shapes in higher dimensions. The expectation that it should translate has come largely through analogy with other geometrical rules that have proved similarly elastic (see "Edges, corners and faces", page 77). When it comes to guaranteeing short routes between points on the surface of a 4D, 5D or 1000D shape, Hirsch's rule has remained one of those niggling unsolved problems of mathematics – a mere conjecture.

How is this relevant? Because, for today's mathematicians, dimensions are not just ›

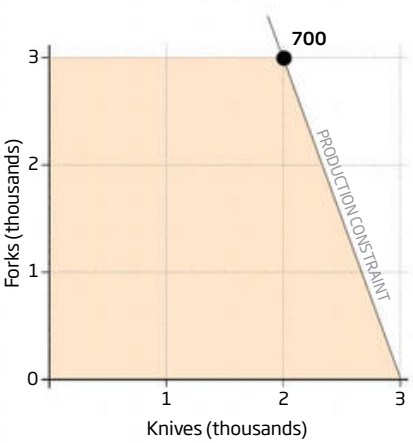
Room for improvement

Many business problems can be reduced to patterns in geometry – as this simple 2D example shows

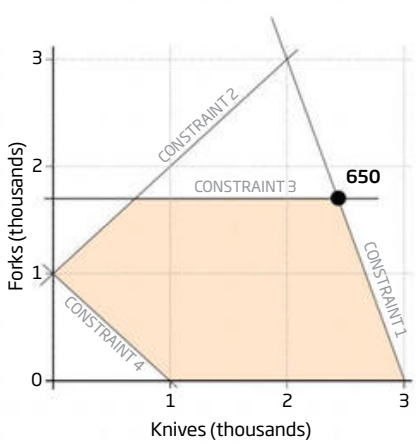
A cutlery factory makes $200 profit for every 1000 knives and $100 for every 1000 forks. With no constraints on production, **more profit** is made by making more of both



In the real world, finite staff or machine resources will mean the more forks you make, the fewer knives you can make. That constrains your operating space and the **maximum profit** you can make



Further constraints, such as demand for cutlery, restrict your operating space to a 2D geometric shape – and the **maximum achievable profit** always lies at a corner of that shape



about space. True, the concept arose because we have three coordinates of location that can vary independently: up-down, left-right and forwards-backwards. Throw in time, and you have a fourth "dimension" that works very similarly, apart from the inexplicable fact that we can move through it in only one direction.

But beyond motion, we often encounter real-world situations where we can vary many more than four things independently. Suppose, for instance, you are making a sandwich for lunch. Your fridge contains 10 ingredients that can be used in varying quantities: cheese, chutney, tuna, tomatoes, eggs, butter, mustard, mayonnaise, lettuce, hummus. These ingredients are nothing other than the dimensions of a sandwich-making problem. This can be treated geometrically: combine your choice of ingredients in any particular way, and your completed snack is represented by a single point in a 10-dimensional space.

Brutish problems

In this multidimensional space, we are unlikely to have unlimited freedom of movement. There might be only two mouldering hunks of cheese in the fridge, for instance, or the merest of scrapings at the bottom of the mayonnaise jar. Our personal preferences might supply other, more subtle constraints to our sandwich-making problem: an eye on the calories, perhaps, or a desire not to mix tuna and hummus. Each of these constraints represents a boundary to our multidimensional space beyond which we cannot move. Our resources and preferences in effect construct a multidimensional polytope through which we must navigate towards our perfect sandwich.

In reality, the decision-making processes in our sandwich-making are liable to be a little haphazard; with just a few variables to consider, and mere gastric satisfaction riding on the outcome, that's not such a problem. But in business, government and science, similar optimisation problems crop up everywhere and quickly morph into brutes with many thousands or even millions of variables and constraints. A fruit importer might have a 1000-dimensional problem to deal with, for instance, shipping bananas from five distribution centres that store varying numbers of fruit to 200 shops with different order sizes. How many items of fruit should be sent from which centres to which shops while minimising total transport costs?

A fund manager might similarly want to arrange a portfolio optimally to balance risk and expected return over a range of stocks; a railway timetabler to decide how best to roster staff and trains; or a factory or hospital manager to work out how to juggle finite machine resources or ward space. Each such problem can be depicted as a geometrical shape whose number of dimensions is the number of variables in the problem, and whose boundaries are delineated by whatever constraints there are (see diagram, left). In each case, we need to box our way through this polytope towards its optimal point.

This is the job of the algorithm.

Its full name is the simplex algorithm, and it emerged in the late 1940s from the work of the US mathematician George Dantzig, who had spent the second world war investigating ways to increase the logistical efficiency of the US air force. Dantzig was a pioneer in the field of what he called linear programming, which

PLAINPICTURE/GOZOOMA

Some heavy-duty mathematics underlies the business of business

uses the mathematics of multidimensional polytopes to solve optimisation problems. One of the first insights he arrived at was that the optimum value of the "target function" – the thing we want to maximise or minimise, be that profit, travelling time or whatever – is guaranteed to lie at one of the corners of the polytope. This instantly makes things much more tractable: there are infinitely many points within any polytope, but only ever a finite number of corners.

If we have just a few dimensions and constraints to play with, this fact is all we need. We can feel our way along the edges of the polytope, testing the value of the target function at every corner until we find its sweet spot. But things rapidly escalate. Even just a 10-dimensional problem with 50 constraints – perhaps trying to assign a schedule of work to 10 people with different expertise and time constraints – may already land us with several billion corners to try out.

The simplex algorithm finds a quicker way through. Rather than randomly wandering along a polytope's edges, it implements a "pivot rule" at each corner. Subtly different variations of this pivot rule exist in different implementations of the algorithm, but often it involves picking the edge along which the target function descends most steeply, thus ensuring each step takes us nearer the optimal value. When a corner is found where no further descent is possible, we know we have arrived at the optimal point.

"Probably tens or hundreds of thousands of calls are made of the simplex algorithm every minute"

Practical experience shows that the simplex method is generally a very slick problem-solver indeed, typically reaching an optimum solution after a number of pivots comparable to the number of dimensions in the problem. That means a likely maximum of a few hundred steps to solve a 50-dimensional problem, rather than billions with a suck-it-and-see approach. Such a running time is said to be "polynomial" or simply "P", the benchmark for practical algorithms that have to run on finite processors in the real world (see "The hardest problem", page 68).

Dantzig's algorithm saw its first commercial application in 1952, when Abraham Charnes and William Cooper at what is now Carnegie Mellon University in Pittsburgh, Pennsylvania, teamed up with Robert Mellon at the Gulf Oil Company to discover how best to blend available stocks of four different petroleum products into an aviation fuel with an optimal octane level. Since then the simplex algorithm has steadily conquered the world, embedded both in commercial optimisation packages and bespoke software products. Wherever anyone is trying to solve a large-scale optimisation problem, the chances are that some computer chip is humming away to its tune. "Probably tens or hundreds of thousands of calls of the simplex method are made every minute," says Jacek Gondzio, an optimisation specialist at the University of Edinburgh, UK.

Yet even as its popularity grew in the 1950s and 1960s, the algorithm's underpinnings were beginning to show signs of strain. To start with, its running time is polynomial only on average. In 1972, US mathematicians Victor Klee and George Minty reinforced this point by running the algorithm around some ingeniously deformed multidimensional "hypercubes". Just as a square has four corners, and a cube eight, a hypercube in n dimensions has 2^n corners. The wonky way Klee and Minty put their hypercubes together meant that the simplex algorithm had to run through all of these corners before landing on the optimal one. In just 41 dimensions, that leaves the algorithm with over a trillion edges to traverse.

EDGES, CORNERS AND FACES

Since Plato laid down his stylus, a lot of work has gone into understanding the properties of 3D shapes, or polyhedra. Perhaps the most celebrated result came from the 18th-century mathematician Leonhard Euler. He noted that every polyhedron has a number of edges that is two fewer than the total of its faces and corners. The cube, for example, has six faces and eight corners, a total of 14, while its edges number 12.

The truncated icosahedron, meanwhile, is the familiar pattern of a standard soccer ball. It has 32 faces (12 pentagonal and 20 hexagonal), 60 corners – and 90 edges.

The French mathematician Adrien-Marie Legendre proved this rule in 1794 for every 3D shape that contains no holes and does not cut through itself in any strange way. As geometry started to grow more sophisticated and extend into higher dimensions in the 19th century, it became clear that Euler's relationship didn't stop there: a simple extension to the rule applies to shapes, or polytopes, in any number of dimensions. For a 4D "hypercube", for example, a variant of the formula guarantees that the total number of corners (16) and faces (24) will be equal to number of edges (32) added to the number of 3D "facets" the shape possesses (8).

The rule derived by Warren Hirsch in 1957 about the maximum distance between two corners of a polyhedron was thought to be similarly cast-iron. Whether it truly is turns out to have surprising relevance to the smooth workings of the modern world.

2000 YEARS OF ALGORITHMS

George Dantzig's simplex algorithm has a claim to be the world's most significant. But algorithms go back much further.

c. 300 BC THE EUCLIDEAN ALGORITHM

From Euclid's mathematical primer *Elements*, this is the grandaddy of all algorithms, showing how, given two numbers, you can find the largest number that divides into both. It has still not been bettered.

820 THE QUADRATIC ALGORITHM

The word algorithm is derived from the name of the Persian mathematician Al-Khwarizmi. Experienced practitioners today perform his algorithm for solving quadratic equations (those containing an x^2 term) in their heads. For everyone else, modern algebra provides the formula familiar from school.

1936 THE UNIVERSAL TURING MACHINE

The British mathematician Alan Turing equated algorithms with mechanical processes – and found one to mimic all the others, the theoretical template for the programmable computer.

1946 THE MONTE CARLO METHOD

When your problem is just too hard to solve directly, enter the casino of chance. John von Neumann, Stanislaw Ulam and Nicholas Metropolis's Monte Carlo algorithm taught us how to play and win.

1957 THE FORTRAN COMPILER

Programming was a fiddly, laborious job until an IBM team led by John Backus invented the first high-level programming language, Fortran. At the centre is the compiler: the algorithm that converts the programmer's instructions into machine code.

1962 QUICKSORT

Extracting a word from the right place in a dictionary is an easy task; putting all the words in the right order in the first place is not. The British mathematician Tony Hoare provided the recipe, now an essential tool in managing databases of all kinds.

1965 THE FAST FOURIER TRANSFORM

Much digital technology depends on breaking down irregular signals into their pure sine-wave components – making James Cooley and John Tukey's algorithm one of the world's most widely used.

1994 SHOR'S ALGORITHM

Peter Shor at Bell Labs found a new, fast algorithm for splitting a whole number into its constituent primes – but it could only be performed by a quantum computer. If ever implemented on a large scale, it would nullify almost all modern internet security.

1998 PAGERANK

The internet's vast repository of information would be of little use without a way to search it. Stanford University's Sergey Brin and Larry Page found a way to assign a rank to every web page – and the founders of Google have been living off it ever since.

A similar story holds for every variation of the algorithm's pivot rule tried since Dantzig's original design: however well it does in general, it always seems possible to concoct some awkward optimisation problems in which it performs poorly. The good news is that these pathological cases tend not to show up in practical applications – though exactly why this should be so remains unclear. "This behaviour eludes any rigorous mathematical explanation, but it certainly pleases practitioners," says Gondzio.

Flashy pretenders

The fault was still enough to spur on researchers to find an alternative to the simplex method. The principal pretender to the throne came along in the 1970s and 1980s with the discovery of "interior point methods", flashy algorithms that, rather than feeling their way around a polytope's surface, drill a path through its core. They came up with a genuine mathematical seal of approval – a guarantee always to run in polynomial time – and typically took fewer steps to reach the optimum point than the simplex method, rarely needing over 100 moves regardless of how many dimensions the problem had.

The discovery generated a lot of excitement, and for a while it seemed that the demise of Dantzig's algorithm was on the cards. Yet it survived and even prospered. The trouble with interior point methods is that each step entails far more computation than a simplex pivot: instead of comparing a target function along a small number of edges, you must analyse all the possible directions within the polytope's interior, a gigantic undertaking. For some huge industrial problems, this trade-off is worth it, but for by no means all. Gondzio estimates that between 80 and 90 per cent of today's linear optimisation problems are still solved by some variant of the simplex algorithm. The same goes for a good few of the even more complex non-linear problems (see "Straight down the line", right). "As a devoted interior-point researcher I have a huge respect for the simplex method," says Gondzio. "I'm doing my best trying to compete."

We would still dearly love to find something better: some new variant of the simplex algorithm that preserves all its advantages, but also invariably runs in polynomial time. For US mathematician and Fields medallist Steve Smale, writing in 1998, discovering such a "strongly polynomial" algorithm was one of 18 outstanding mathematical questions to be dealt with in the 21st century.

Yet finding such an algorithm may not now even be possible.

That is because the existence of such an improved, infallible algorithm depends on a more fundamental geometrical assumption – that a short enough path around the surface of a polytope between two corners actually exists. Yes, you've got it: the Hirsch conjecture.

The fates of the conjecture and the algorithm have always been intertwined.

SIMON GARDINER

"Cases where the algorithm fails have tended not to show up in practice – a pleasing behaviour that eludes explanation"

STRAIGHT DOWN THE LINE

In 1948, a young and nervous George Dantzig was presenting at a conference of eminent economists and statisticians in Wisconsin. As he spoke about his new simplex algorithm, a rather large hand was raised in objection at the back of the room. It was that of the renowned mathematician Harold Hotelling. "But we all know the world is non-linear," he said.

It was a devastating put-down. The simplex algorithm's success in solving optimisation problems depends on assuming that variables change in response to other variables along nice straight lines. A cutlery company increasing its expenditure on metal, for example, will produce proportionately more finished knives, forks and profit the next month.

In fact, as Hotelling pointed out, the real world is jam-packed with non-linearity. As the cutlery company expands, economies of scale may mean the marginal cost of each knife or fork drops, making for a non-linear profit boost. In geometrical terms, such problems are represented by multidimensional shapes just as linear problems are, but ones bounded by curved faces that the simplex algorithm should have difficulty crawling round.

Surprisingly, though, linear approximations to non-linear processes turn out to be good enough for most practical purposes. "I would guess that 90 or 95 per cent of all optimisation problems solved in the world are linear programs," says Jacek Gondzio of the University of Edinburgh, UK. For those few remaining problems that do not submit to linear wiles, there is a related field of non-linear programming – and here too, specially adapted versions of the simplex algorithm have come to play an important part.

Hirsch was himself a pioneer in operational research and an early collaborator of Dantzig's, and it was in a letter to Dantzig in 1957 musing about the efficiency of the algorithm that Hirsch first formulated his conjecture.

Until recently, little had happened to cast doubt on it. Klee proved it true for all 3D polyhedra in 1966, but had a hunch the same did not hold for higher-dimensional polytopes. In his later years, he made a habit of suggesting it as a problem to every freshly scrubbed researcher he ran across. In 2001 one of them, a young Spaniard called Francisco Santos, now at the University of Cantabria in Santander, took on the challenge.

As is the way of such puzzles, it took time. After almost a decade working on the problem, Santos was ready to announce his findings at a conference in Seattle in 2010, and he published a paper detailing his findings in 2012. In it, he describes a 43-dimensional polytope with 86 faces. According to Hirsch's conjecture, the longest path across this shape would have (86 - 43) steps, that is, 43 steps. But Santos was able to establish conclusively that it contains a pair of corners at least 44 steps apart.

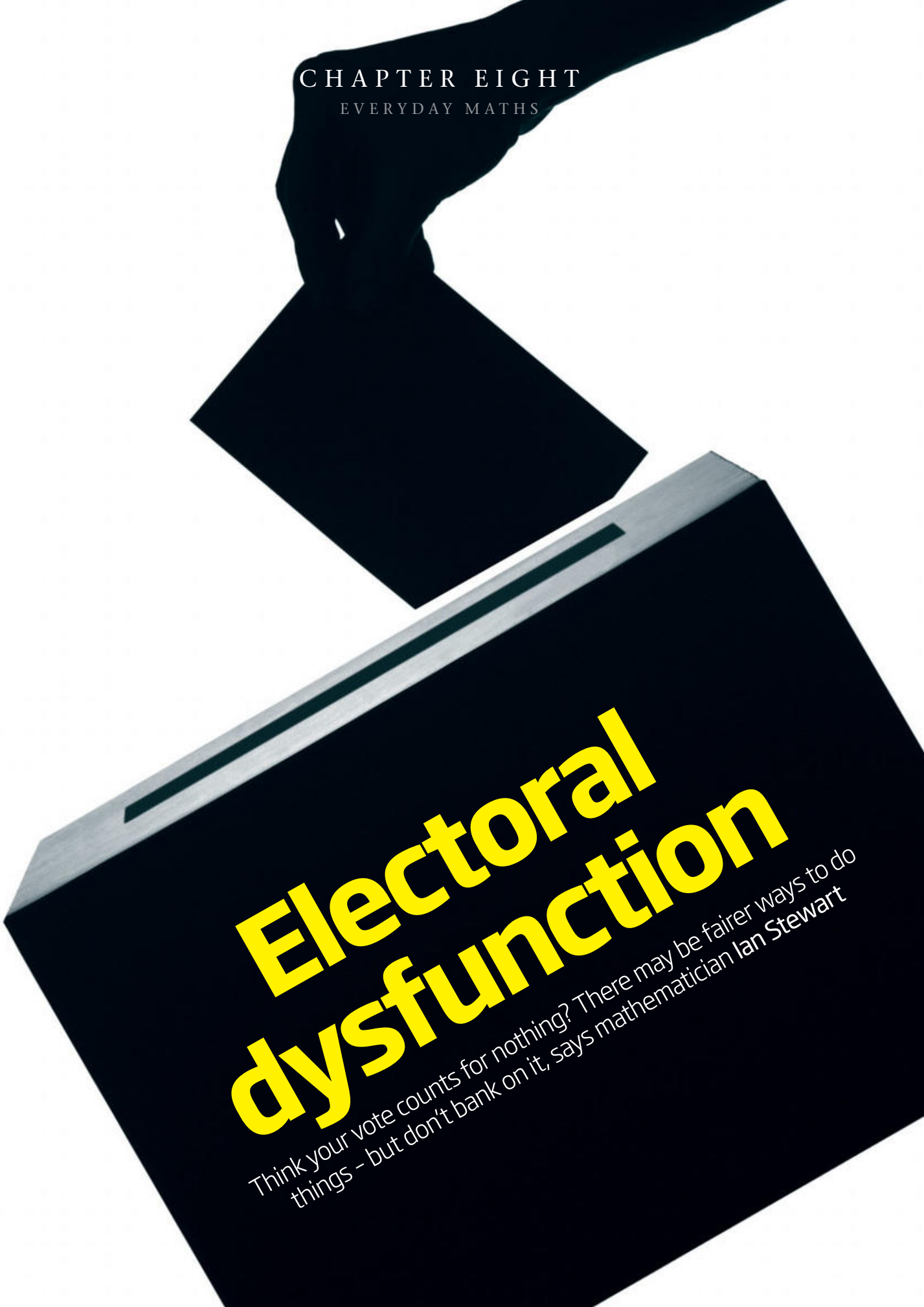
If only for a single special case, Hirsch's conjecture had been proved false. "It settled a problem that we did not know how to approach for many decades," says Gil Kalai of the Hebrew University of Jerusalem. "The entire proof is deep, complicated and very elegant. It is a great result."

A great result, true, but decidedly bad news for the simplex algorithm. Since Santos's first disproof, further Hirsch-defying polytopes have been found in dimensions as low as 20. The only known limit on the shortest distance between two points on a polytope's surface is now contained in a mathematical expression derived by Kalai and Daniel Kleitman of the Massachusetts Institute of Technology in 1992. This bound is much larger than the one the Hirsch conjecture would have provided, had it proved to be true. It is far too big, in fact, to guarantee a reasonable running time for the simplex method, whatever fancy new pivot rule we might dream up. If this is the best we can do, it may be that Smale's goal of an idealised algorithm will remain forever out of reach – with potentially serious consequences for the future of optimisation.

All is not lost, however. A highly efficient variant of the simplex algorithm may still be possible if the so-called polynomial Hirsch conjecture is true. This would considerably tighten Kalai and Kleitman's bound, guaranteeing that no polytopes have paths disproportionately long compared with their dimension and number of faces. A topic of interest before the plain-vanilla Hirsch conjecture melted away, the polynomial version has been attracting intense attention since Santos's announcement, both as a deep geometrical conundrum and a promising place to sniff around for an optimally efficient optimisation procedure.

As yet, there is no conclusive sign that the polynomial conjecture can be proved either. "I am not confident at all," says Kalai. Not that this puts him off. "What is exciting about this problem is that we do not know the answer."

A lot could be riding on that answer. As the algorithm continues to hum away in those basements it is still, for the most part, telling us what we want to know in the time we want to know it. But its own fate is now more than ever in the hands of the mathematicians. ■



Electoral dysfunction

Think your vote counts for nothing? There may be fairer ways to do things – but don't bank on it, says mathematician **Ian Stewart**

IN AN ideal world, elections should be two things: free and fair. Every adult, with a few sensible exceptions, should be able to vote for a candidate of their choice, and each single vote should be worth the same.

Ensuring a free vote is a matter for the law. Making elections fair is more a matter for mathematicians. They have been studying voting systems for hundreds of years, looking for sources of bias that distort the value of individual votes, and ways to avoid them. Along the way, they have turned up many paradoxes and surprises. What they have not done is come up with the answer. With good reason: it probably doesn't exist.

The many democratic electoral systems in use around the world attempt to strike a balance between mathematical fairness and political considerations such as accountability and the need for strong, stable government. Take first-past-the-post or "plurality" voting, which used for national elections in Canada, India, the US and the UK. Its principle is simple: each electoral division elects one representative, the candidate who gained the most votes.

This system scores well on stability and accountability, but in terms of mathematical fairness it is a dud. Votes for anyone other than the winning candidate are disregarded. If more than two parties with substantial support contest a constituency, as is typical in Canada, India and the UK, a candidate does not have to get anything like 50 per cent of the votes to win, so a majority of votes are "lost".

Dividing a nation or city into bite-sized chunks for an election is itself a fraught business (see "Borderline case", right) that invites other distortions, too. A party can win by being only just ahead of its competitors in most electoral divisions. In the UK general election in 2005, for example, the ruling Labour party won 55 per cent of the seats on just 35 per cent of the total votes. If a candidate or party is slightly ahead in a bare majority of electoral divisions but a long way behind in others, they can win even if a competitor gets more votes overall – as happened most notoriously in recent history with Donald Trump's victory in the US presidential election of 2016.

The anomalies of a plurality voting system can be more subtle, though, as mathematician Donald Saari at the University of California, Irvine, showed. Suppose 15 people are asked to rank their liking for milk (M), beer (B), or wine (W). Six rank them M-W-B, five B-W-M, and four W-B-M. In a plurality system where only first preferences count, the outcome is simple: milk wins with 40 per cent of the vote, followed by beer, with wine trailing in last.

So do voters actually prefer milk? Not a bit of it. Nine voters prefer beer to milk, and nine prefer wine to milk – clear majorities in both cases. Meanwhile, 10 people prefer wine to beer. By pairing off all these preferences, we see the truly preferred order to be W-B-M – the exact reverse of what the voting system produced. In fact Saari showed that given a set of voter preferences you can design a system that produces any result you desire.

In the example above, simple plurality voting produced an anomalous outcome because the alcohol drinkers stuck together: wine and beer drinkers both nominated the other as their second preference and gave milk a big thumbs-down. Similar things happen in politics when two parties appeal to the same kind of voters, splitting their votes between them and allowing a third party unpopular with the majority to win the election.

Can we avoid that kind of unfairness while keeping the advantages of a first-past-the-post system? Only to an extent. One possibility is a second "run-off" election between the two top-ranked candidates, as happens in France and in many presidential elections elsewhere. But there is no guarantee that the two candidates with the widest potential support even make the run-off. In the 2002 French presidential election, for example, so many left-wing candidates stood in the first round that all of them were eliminated, leaving two right-wing candidates, Jacques Chirac and Jean-Marie Le Pen, to contest the run-off.

Order, order

Another strategy allows voters to place candidates in order of preference, with a 1, 2, 3 and so on. After the first-preference votes have been counted, the candidate with the lowest score is eliminated and the votes reapportioned to the next-choice candidates on those ballot papers. This process goes on until one candidate has the support of over 50 per cent of the voters. This system, called the instant run-off or alternative or preferential vote, is used in elections to the Australian House of Representatives, as well as in several US cities. A move to introduce this system for UK parliamentary elections was defeated in a referendum in 2011.

Preferential voting comes closer to being ›

COMSTOCK/GETTY

BORDERLINE CASE

In first-past-the-post or "plurality" systems, borders matter. To ensure that each vote has roughly the same weight, each constituency should have roughly the same number of voters. Threading boundaries between and through centres of population on the pretext of ensuring fairness is also a great way to cheat for your own benefit – a practice known as gerrymandering, after a 19th-century governor of Massachusetts, Elbridge Gerry, who created an electoral division whose shape reminded a local newspaper editor of a salamander.

Suppose a city controlled by the Liberal Republican (LR) party has a voting population of 900,000 divided into three constituencies. Polls show that at the next election LR is heading for defeat – 400,000 people intend to vote for it but the 500,000 others will opt for the Democratic Conservative (DC) party. If the boundaries were to keep the proportions the same, each constituency would contain roughly 130,000 LR voters and 170,000 DC voters, and DC would take all three seats – the usual inequity of a plurality voting system.

In reality, voters inclined to vote for one party or the other will probably clump together in the same neighbourhoods of the city, so LR might well retain one seat. However, it could be all too easy for LR to redraw the boundaries to reverse the result and secure itself a majority – as the following two dividing strategies show.

Each square represents 100,000 voters

Liberal Republicans (**LR**): Total votes 400,000
Democratic Conservatives (**DC**): Total votes 500,000

SCENARIO 1

Constituency 1: LR 100,000, **DC 200,000**
Constituency 2: **LR 200,000**, DC 100,000
Constituency 3: LR 100,000, **DC 200,000**

LR 1 SEAT, DC 2 SEATS – **DC wins**

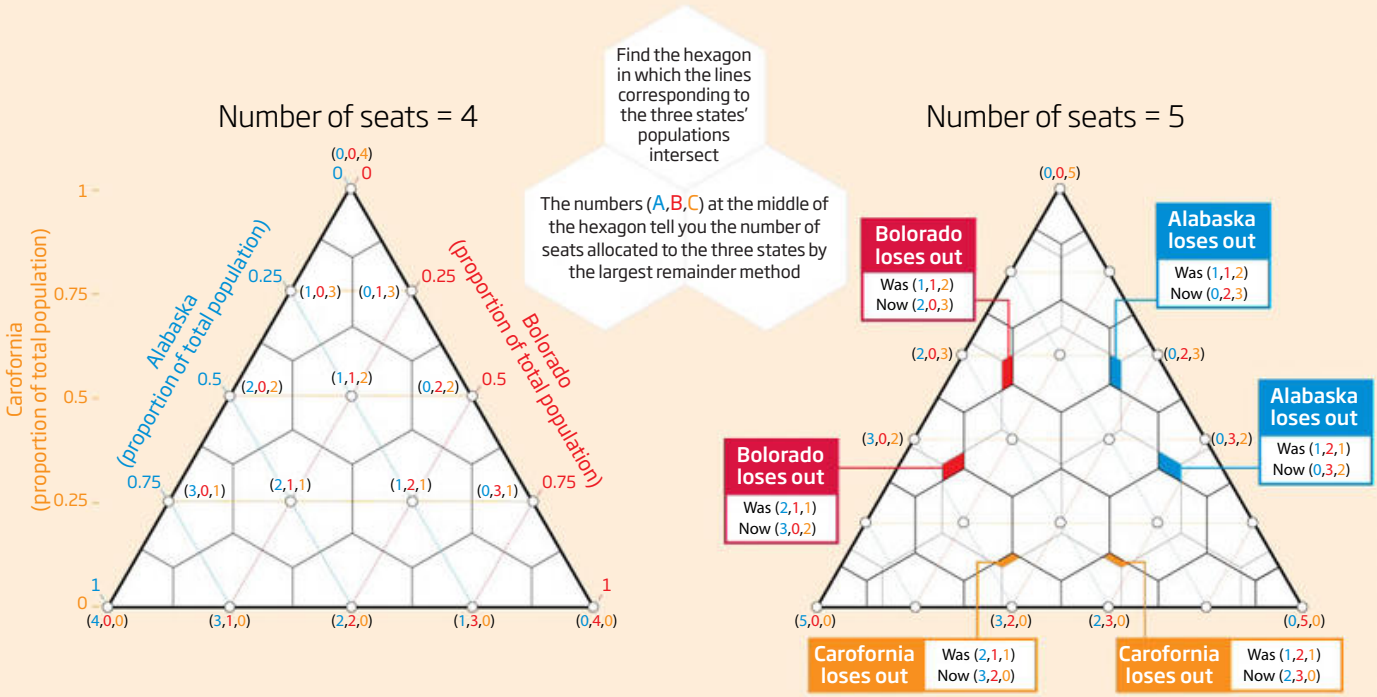
SCENARIO 2

Constituency 1 (left): **LR 200,000**, DC 100,000
Constituency 2 (top): LR no votes, **DC 300,000**
Constituency 3 (bottom): **LR 200,000**, DC 100,000

LR 2 SEATS, DC 1 SEAT – **LR wins**

PROPORTIONAL PARADOX

A state can lose representation if the number of seats in a national parliament increases, even if its population stays the same



Although elections to the US House of Representatives use a first-past-the-post voting system, the constitution requires that seats be "apportioned among the several states according to their respective numbers" – that is, divvied up proportionally. In 1880, the chief clerk of the US Census Bureau, Charles Seaton, discovered that Alabama would get eight seats in a 299-seat House, but only seven in a 300-seat House.

This "Alabama paradox" was caused by an algorithm known as the largest remainder method, which was used to round the number of seats a state would receive under strict proportionality to a whole number.

Suppose for simplicity's sake that a nation of 39 million voters has a parliament with four seats – giving a quota of 9.75 million voters per seat. The seats must, however, be shared among three states, Alabaska, Bolorado and Carofornia, with voting populations of 21, 13 and 5 million, respectively. Dividing these numbers by the quota gives each state's fair proportion of seats. Rounded down to an integer, this number of seats is given to the states. Any seats left over go to the state or states with the highest remainders.

	Alabaska	Bolorado	Carofornia
Fair proportion	2.15	1.33	0.51
Rounded-down integer	2	1	0
Remainder	0.15	0.33	0.51
Extra seats	0	0	1
Total seats	2	1	1

The rounded-down integers allocate three seats. The fourth goes to Carofornia, the state with the largest remainder.

Suppose now the number of seats increases from four to five. The quota is 39 million divided by 5, or 7.8 million, and so our table becomes:

	Alabaska	Bolorado	Carofornia
Fair proportion	2.69	1.67	0.64
Rounded-down integer	2	1	0
Remainder	0.69	0.67	0.64
Extra seats	1	1	0
Total seats	3	2	0

The rounded-down integers account for three seats as before. The two spare go to Alabaska and Bolorado, which have the two largest remainders, and Carofornia loses its only seat. (The US Constitution stipulates that each state must have at least one representative, which would protect Carofornia in this case – the size of the House would have to be increased by one seat.)

The precise conditions that lead to the Alabama paradox are mathematically complex. For three states they can be portrayed graphically, as above. The left-hand diagram shows the populations (as a fraction of the country's total) and fair proportions of three states in the case of four seats; the right-hand side superimposes the diagram for five seats. The Alabama paradox occurs for the shaded population combinations: our example lies in the leftmost orange-shaded region.

Such quirks mean that seats in proportional systems are now generally apportioned using algorithms known as divisor methods. These work by dividing voting populations by a common factor so that when the fair proportions are rounded to a whole number, they add up to the number of available seats. But this method is not foolproof: it sometimes gives a constituency more seats than the whole number closest to its fair proportion.

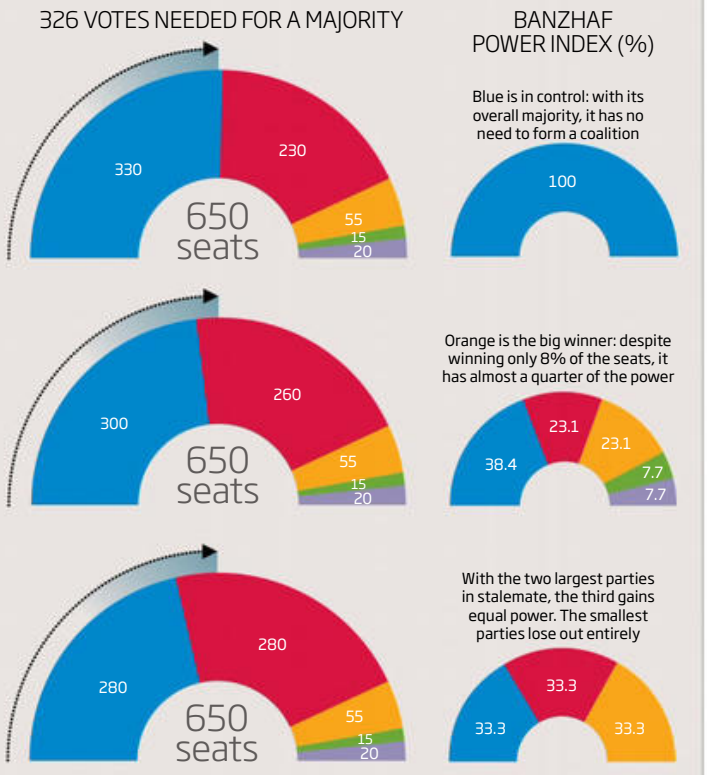
POWER IN THE BALANCE

One criticism of proportional voting systems is that they make it less likely that one party wins a majority of the seats available, thus increasing the power of smaller parties as "king-makers" who can swing the balance between rival parties as they see fit. The same can happen in a plurality system if the electoral arithmetic delivers a hung parliament, in which no party has an overall majority – as might happen in the UK after its election next week.

Where does the power reside in such situations? One way to quantify that question is the Banzhaf power index. First, list all combinations of parties that could form a majority coalition, and in all of those coalitions count how many times a party is a "swing" partner that could destroy the majority if it dropped out. Dividing this number by the total number of swing partners in all possible majority coalitions gives a party's power index.

For example, imagine a parliament of six seats in which party A has three seats, party B has two and party C has one. There are three ways to make a coalition with a majority of at least four votes: AB, AC and ABC. In the first two instances, both partners are swing partners. In the third instance, only A is – if either B or C dropped out, the remaining coalition would still have a majority. Among the total of five swing partners in the three coalitions, A crops up three times and B and C once each. So A has a power index of 3 ÷ 5, or 0.6, or 60 per cent – more than the 50 per cent of the seats it holds – and B and C are each "worth" just 20 per cent.

In a realistic situation, the calculations are more involved. The diagram on the right shows how the power shifts dramatically when there is no majority in a hypothetical parliament of 650 seats in which five voting blocs are represented.



The influence of small parties grows in a hung parliament even if their own vote does not

fair than plurality voting, but it does not eliminate ordering paradoxes. The Marquis de Condorcet, a French mathematician, noted this as early as 1785. Suppose we have three candidates, A, B and C, and three voters who rank them A-B-C, B-C-A and C-A-B. Voters prefer A to B by 2 to 1. But B is preferred to C and C preferred to A by the same margin of 2 to 1. To quote the Dodo in *Alice in Wonderland*: "Everybody has won and all must have prizes."

One type of voting system avoids such circular paradoxes entirely: proportional representation. Here a party is awarded a number of parliamentary seats in direct proportion to the number of people who voted for it. Such a system is undoubtedly fairer in a mathematical sense than either plurality or preferential voting, but it has political drawbacks. It implies large, multi-representative constituencies; the best shot at truly proportional representation comes with just one constituency, the system used in Israel. But large constituencies weaken the link between voters and their representatives. Candidates are often chosen from a centrally determined list, so voters have little or no control over who represents them. What's more, proportional systems tend to produce coalitions of two or more parties, potentially leading to unstable and ineffectual government – although plurality systems are not immune to such problems, either (see "Power in the balance", above).

Proportional representation has its own mathematical wrinkles. There is no way, for example, to allocate a whole number of seats in exact proportion to a larger population. This can lead to an odd situation in which increasing the total number of seats available reduces the representation of an individual constituency, even if its population stays the same (see "Proportional paradox", left).

> "No one voting system satisfies all conditions of fairness"

Such imperfections led the American economist Kenneth Arrow to list in 1963 the general attributes of an idealised fair voting system. He suggested that voters should be able to express a complete set of their preferences; no single voter should be allowed to dictate the outcome of the election; if every voter prefers one candidate to another, the final ranking should reflect that; and if a voter prefers one candidate to a second, introducing a third candidate should not reverse that preference.

All very sensible. There's just one problem: Arrow and others went on to prove that no conceivable voting system could satisfy all four conditions. In particular, there will always be the possibility that one voter, simply by changing their vote, can change the overall preference of the whole electorate.

So we are left to make the best of a bad job. Some less fair systems produce governments with enough power to actually do things, though most voters may disapprove; some fairer systems spread power so thinly that any attempt at government descends into partisan infighting. Crunching the numbers can help, but deciding which is the lesser of the two evils is ultimately a matter not for mathematics, but for human judgement. ■

As easy as **pie**

Sharing a pizza with a friend but not sure you are getting fair shares? **Stephen Ornes** discovers there's a guaranteed way to find out

LUNCH with a colleague from work should be a time to unwind – the most taxing task being to decide what to eat, drink and choose for dessert. For Rick Mabry and Paul Deiermann it has never been that simple. They can't think about sharing a pizza, for example, without falling headlong into the mathematics of how to slice it up. "We went to lunch together at least once a week," says Mabry, recalling the early 1990s when they were both at Louisiana State University, Shreveport. "One of us would bring a notebook, and we'd draw pictures while our food was getting cold."

The problem that bothered them was this. Suppose the harried waiter cuts the pizza off-centre, but with all the edge-to-edge cuts crossing at a single point, and with the same angle between adjacent cuts. The off-centre cuts mean the slices will not all be the same size, so if two people take turns to take neighbouring slices, will they get equal shares by the time they have gone right round the pizza – and if not, who will get more?

Of course you could estimate the area of each slice, tot them all up and work out each person's total from that. But these guys are mathematicians, and so that wouldn't quite do. They wanted to be able to distil the problem down to a few general, provable rules that avoid exact calculations, and that work every time for any circular pizza.

As with many mathematical conundrums, the answer has arrived in stages – each looking at different possible cases of the problem. The easiest example to consider is when at least one cut passes plumb through the centre of the pizza. A quick sketch shows that the pieces then pair up on either side of the cut through the centre, and so can be divided evenly between the two diners, no matter how many cuts there are.

So far so good, but what if none of the cuts passes through the centre? For a pizza cut once, the answer is obvious by inspection: whoever eats the centre eats more. The case of a pizza cut twice, yielding four slices, shows the same result: the person who eats the slice that contains the centre gets the bigger portion. That turns out to be an anomaly to the three general rules that deal with greater numbers of cuts, which would emerge over subsequent years to form the complete pizza theorem.

The first proposes that if you cut a pizza through the chosen point with an even number of cuts more than 2, the pizza will be divided evenly between two diners who each take alternate slices. This side of the problem was first explored in 1967 by one L. J. Upton in *Mathematics Magazine*. Upton didn't bother with two cuts: he asked readers to prove that in the case of four cuts (making eight slices) the diners can share the pizza equally. Next came the general solution for an even number of cuts greater than 4, which first turned up as an answer to Upton's challenge in 1968, with elementary algebraic calculations of the exact area of the different slices revealing that, again, the pizza is always divided equally between the two diners.

"Most mathematicians would have thought, 'I'm not going to look at it.' We were stupid enough to try"

With an odd number of cuts, things start to get more complicated. Here the pizza theorem says that if you cut the pizza with 3, 7, 11, 15… cuts, and no cut goes through the centre, then the person who gets the slice that includes the centre of the pizza eats more in total. If you use 5, 9, 13, 17… cuts, the person who gets the centre ends up with less (see "How to cut a pizza", page 86).

Rigorously proving this to be true, however, has been a tough nut to crack. So difficult, in fact, that Mabry and Deiermann have only just finalised a proof that covers all possible cases.

Their quest started in 1994, when Deiermann showed Mabry a revised version of the pizza problem, again published in *Mathematics Magazine*. Readers were invited to prove two specific cases of the pizza theorem. First, that if a pizza is cut three times (into six slices), the person who eats the slice containing the pizza's centre eats more. Second, that if the pizza is cut five times (making 10 slices), the opposite is true and the person who eats the centre eats less.

The first statement was posed as a teaser: it had already been proved by the authors. The second statement, however, was preceded by an asterisk – a tiny symbol which, in *Mathematics Magazine*, can mean big trouble. It indicates that the proposers haven't yet proved the proposition themselves. "Perhaps most mathematicians would have thought, 'If those guys can't solve it, I'm not going to look at it.'" Mabry says. "We were stupid enough to look at it."

Deiermann quickly sketched a solution to the three-cut problem – "one of the most clever things I've ever seen," as Mabry recalls. The pair went on to prove the statement for five cuts – even though new tangles emerged in the process – and then proved that if you cut the pizza seven times, you get the same result as for three cuts: the person who eats the centre of the pizza ends up with more.

Boosted by their success, they thought they might have stumbled across a technique that could prove the entire pizza theorem once and for all. For an odd number of cuts, opposing slices inevitably go to different diners, so an intuitive solution is to simply compare the sizes of opposing slices and figure out who gets more, and by how much, before moving on to the next pair. Working your way around the pizza pan, you tot up the differences and ›

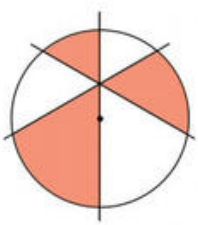
DAVIES AND STARR/DIGITAL VISION

How to cut a pizza

The pizza conjecture asks who will get the bigger portion of a pizza cut off-centre, assuming the diners (A and B) take alternate slices and that the angles between adjacent cuts are all equal

A's slices B's slices • Pizza centre *n* Number of cuts

As long as one cut goes through the centre, both diners get an equal amount of pizza, assuming they choose alternate slices

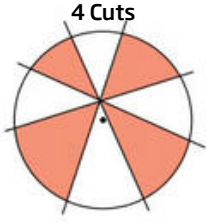


EQUAL AMOUNTS OF PIZZA

Problems start when the cuts don't go through the centre. Working out who gets the most depends on the number of cuts made in the pizza

THE PIZZA CONJECTURE

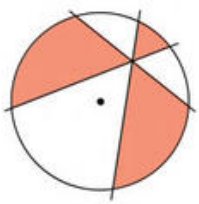
4 Cuts



EQUAL AMOUNTS OF PIZZA

(also for any even $n > 4$)

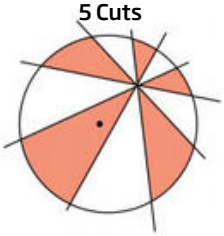
3 Cuts



B EATS MORE WHEN B GETS SLICE CONTAINING CENTRE OF PIZZA

(also for $n = 7,11,15...$)

5 Cuts

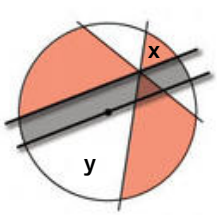


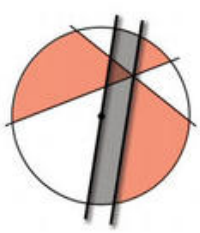
B EATS MORE WHEN A GETS SLICE CONTAINING CENTRE OF PIZZA

(also for $n = 9,13,17...$)

PIZZA PROOF

Rick Mabry and Paul Deiermann found a way to prove the pizza conjecture that involves comparing opposite slices in turn





Instead of looking at the actual slices (x and y, say), they drew a line parallel to each cut running through the centre of the pizza

They then used the "rectangular" shaded areas as a measure of the difference in area of opposing slices. Plug that in to some complicated algebra and the proof arises

there's your answer.

Simple enough in principle, but it turned out to be horribly difficult in practice to come up with a solution that covered all the possible numbers of odd cuts. Mabry and Deiermann hoped they might be able to deploy a deft geometrical trick to simplify the problem. The key was the area of the rectangular strips lying between each cut and a parallel line passing through the centre of the pizza (see diagram). That's because the difference in area between two opposing slices can be easily expressed in terms of the areas of the rectangular strips defined by the cuts. "The formula for [the area of] strips is easier than for slices," Mabry says. "And the strips give some very nice visual proofs of certain aspects of the problem."

Unfortunately, the solution still included a complicated set of sums of algebraic series involving tricky powers of trigonometric functions. The expression was ugly, and even though Mabry and Deiermann didn't have to calculate the total exactly, they still had to prove it was positive or negative to find out who gets the bigger portion. It turned out to be a massive hurdle. "It ultimately took 11 years to figure that out," says Mabry.

"There are a host of other pizza problems – who gets more crust, for example, and who gets most cheese"

Over the following years, the pair returned occasionally to the pizza problem, but with only limited success. The breakthrough came in 2006, when Mabry was on a vacation in Kempten im Allgäu in the far south of Germany. "I had a nice hotel room, a nice cool environment, and no computer," he says. "I started thinking about it again, and that's when it all started working." Mabry and Deiermann – who by now was at Southeast Missouri State University in Cape Girardeau – had been using computer programs to test their results, but it wasn't until Mabry put the technology aside that he saw the problem clearly. He managed to refashion the algebra into a manageable, more elegant form.

Back home, he put computer technology to work again. He suspected that someone, somewhere must already have worked out the simple-looking sums at the heart of the new expression, so he trawled the online world for theorems in the vast field of combinatorics – an area of pure mathematics concerned with listing, counting and rearranging – that might provide the key result he was looking for.

Eventually he found what he was after: a 1999 paper that referenced a mathematical statement from 1979. There, Mabry found the tools he and Deiermann needed to show whether the complex algebra of the rectangular strips came out positive or negative. The rest of the proof then fell into place.

So, with the pizza theorem proved, will all kinds of important practical problems now be easier to deal with? In fact there don't seem to be any such applications – not that Mabry is unduly upset. "It's a funny thing about some mathematicians," he says. "We often don't care if the results have applications because the results are themselves so pretty." Sometimes these solutions to abstract mathematical problems do show their face in unexpected places. For example, a 19th-century mathematical curiosity called the "space-filling curve" – a sort of early fractal curve – recently resurfaced as a model for the shape of the human genome.

Mabry and Deiermann have gone on to examine a host of other pizza-related problems. Who gets more crust, for example, and who will eat the most cheese? And what happens if the pizza is square? Equally appetising to the mathematical mind is the question of what happens if you add extra dimensions to the pizza. A three-dimensional pizza, one might argue, is a calzone – a bread pocket filled with pizza toppings – suggesting a whole host of calzone conjectures, many of which Mabry and Deiermann have already proved. It's a passion that has become increasingly theoretical over the years. So if on your next trip to a pizza joint you see someone scribbling formulae on a napkin, it's probably not Mabry. "This may ruin any pizza endorsements I ever hoped to get," he says, "but I don't eat much American pizza these days." ■

They're either mathematicians or they're feeling lucky

AMERICANA IMAGES/WILDCARD IMAGES

What's luck got to do with it?

Even if you can't beat the system, there are some cunning ways to tilt the odds in your favour. **Helen Thomson** takes a punt

IN 2004, Londoner Ashley Revell sold his house, all his possessions and cashed in his life savings. It raised £76,840. He flew to Las Vegas, headed to the roulette table and put it all on red.

The wheel was spun. The crowd held its breath as the ball slowed, bounced four or five times, and finally settled on number seven. Red seven.

Revell's bet was a straight gamble: double or nothing. But when Edward Thorp, a mathematics student at the Massachussetts Institute of Technology, went to the same casino some 40 years previously, he knew pretty well where the ball was going to land. He walked away with a profit, took it to the racecourse, the basketball court and the stock market, and became a multimillionaire. He wasn't on a lucky streak, he was using his knowledge of mathematics to understand, and beat, the odds.

No one can predict the future, but the powers of probability can help. Armed with this knowledge, a high-school mathematics education and £50, I headed off to find out how Thorp, and others like him, have used mathematics to beat the system. Just how much money could probability make me?

When Thorp stood at the roulette wheel in the summer of 1961 there was no need for nerves – he was armed with the first "wearable" computer, one that could predict the outcome of the spin. Once the ball was in play, Thorp fed the computer information about the speed and position of the ball and the wheel using a microswitch inside his shoe. "It would make a forecast about a probable result, and I'd bet on neighbouring numbers," he says.

Thorp's device would now be illegal in a casino, and in any case getting a computer to do the work wasn't exactly what I had in mind. However, there is a simple and sure-fire way to win at the roulette table – as long as you have deep pockets and a faith in probability theory.

A spin of the roulette wheel is just like the toss of a coin. Each spin is independent, with a 50:50 chance of the ball landing on black or red. Contrary to intuition, a black number is just as likely to appear after a run of 20 consecutive black numbers as the seemingly more likely red. ›

"Go into any casino with normal blackjack rules and you can have a modest advantage without much effort"

This randomness means there is a way of using probability to ensure a profit: always bet on the same colour, and if you lose, double your bet on the next spin. Because your colour will come up eventually, this method will always produce a profit. The downside is that you'll need a big pot of cash to stay in the game: a losing streak can escalate your bets very quickly. Seven unlucky spins on a £10 starting bet will have you parting with a hefty £1280 on the next. Unfortunately, your winnings don't escalate in the same way: when you do win, you will only make a profit equal to your original stake. So while the theory itself is sound, be careful. The roulette wheel is likely to keep on taking your money longer than you can remain solvent.

With that in mind, I turned my back on roulette and followed Thorp into the card game blackjack. In 1962 he published a book called *Beat the Dealer*, which proved what many had long suspected: by keeping track of the cards, you can tip the odds in your favour. He earned thousands of dollars putting his proof into practice.

The method is now known as card counting. So does it still work? Could I learn to do it? And is it legal?

"It's certainly not illegal," Thorp assures me. "The casino can't see inside your head – yet." What's more, after a brief tutorial, it doesn't sound too difficult. "If you went into any casino that had basic blackjack rules, learned the method of card counting that I've taught you, you'd have a modest advantage without much effort," says Thorp.

Basic card counting is simple. Blackjack starts with each player being dealt two cards face up. Face cards count as 10 and the ace as 1 or 11 at the player's discretion. The aim is to have as high a total as possible without "busting" – going over 21. To win, you must achieve a score higher than the dealer's. Cards are dealt from a "shoe" – a box of cards made up of three to six decks. Players can stick with the two cards they are dealt or "hit" and receive an extra card to try to get closer to 21. If the dealer's total is 16 or less, the dealer must hit. At the end of each round, used cards are discarded.

The basic idea of card counting is to keep track of those discarded cards to know what's left in the shoe. That's because a shoe rich in high cards will slightly favour you, while a shoe rich in low cards is slightly better for the dealer. With lots of high cards still to be dealt you are more likely to score 20 or 21 with your first two cards, and the dealer is more likely to bust if his initial cards are less than 17. An abundance of low cards benefits the

You can't beat bookies, but you can play them off against each other

ASH KNOTEK/SNAPPERS RIGHT: CORBIS

dealer for similar reasons.

If you keep track of which cards have been dealt, you can gauge when the game is swinging in your favour. The simplest way is to start at zero and add or subtract according to the dealt cards. Add 1 when low cards (two to six) appear, subtract 1 when high cards (10 or above) appear, and stay put on seven, eight and nine. Then place your bets accordingly – bet small when your running total is low, and when your total is high, bet big. This method can earn you a positive return of up to 5 per cent on your investment, says Thorp.

After a bit of practice at home, I head off to my nearest casino. Trying to blend in among the rich young things, the shady mafia types and the glamorous cocktail waitresses was one thing; counting cards while trying to remain calm was another. "If they suspect that you're counting cards, they'll ask you to move to a different game or throw you out completely," one of the casino's regulars tells me.

After a few hours I begin to get the hang of it, and eventually walk away with a profit of £12.50 on a total stake of £30. The theory is good, but in practice it's a lot of effort for a small return. It would be a lot easier if I could just win the lottery. How can I improve my chances there?

"The lotto strategy is clear: go for numbers above 31, and pick ones that are situated around the edges of the ticket. Then if you do win, you'll win big"

In it to win it

The evening of 14 January 1995 was one that Alex White will never forget. He matched all six numbers on the UK National Lottery, with a massive estimated jackpot of £16 million. Unfortunately, White (not his real name) only won £122,510 because 132 other people also matched all six numbers and took a share of the jackpot.

There are dozens of books that claim to improve your odds of winning the lottery. None of them works. Every combination of numbers has the same odds of winning as any other – 1 in 13,983,816 in the case of the UK's 49-ball "Lotto" game of the time. But, as White's story shows, the fact that you could have to share the jackpot suggests a way to maximise any winnings. Your chances of success may be tiny, but if you win with numbers nobody else has chosen, you win big.

So how do you choose a combination unique to you? You won't find the answer at the National Lottery headquarters – they don't give out any information about the numbers people choose. That didn't stop Simon Cox, a mathematician at the University of Southampton, UK from trying. In the mid 1990s, a few years after the UK National Lottery had begun, Cox worked out UK lottery players' favourite figures by analysing data from 113 lottery draws. He compared the winning numbers with how many people had matched four, five or six of them, and thereby inferred which numbers are most popular.

And what were the magic numbers? Seven was the favourite, chosen 25 per cent more often than the least popular number, 46. Numbers 14 and 18 were also popular, while 44 and 45 were among the least favourite. The most noticeable preference was for numbers up to 31. "They call this the birthday effect," says Cox. "A lot of people use their date of birth."

Several other patterns emerged. The most popular numbers are clustered around the centre of the form people fill in to make their selection, suggesting that players are influenced by its layout. Similarly, thousands of players appear to just draw a diagonal line through a group of numbers on the form. There is also a clear dislike of consecutive numbers. "People refrain from choosing numbers next to each other, even though getting 1, 2, 3, 4, 5, 6 is as likely as any other combination," says Cox. Numerous studies on the US, Swiss and Canadian lotteries have produced similar findings.

To test the idea that picking unpopular numbers can maximise your winnings, Cox simulated a virtual syndicate that bought 75,000 tickets each week, choosing its numbers at random. Using the real results of the first 224 UK lottery draws, he calculated that his syndicate would have won a total of £7.5 million – on an outlay of £16.8 million. If his syndicate had stuck to unpopular numbers, however, it would have more than doubled its winnings.

So the strategy is clear: go for numbers above 31, and pick ones that are clumped together or situated around the edges of the form. Then if you match all the numbers, you won't have to share with dozens of others. Unfortunately, probability also predicts that you won't match all the numbers in a weekly draw for a good few centuries. Perhaps I'm best off heading for the bookmaker.

Although it would be nearly impossible to beat a seasoned bookie at his own game, play two or three bookies against each other and you can come up a winner. So claims John Barrow, professor of mathematics at the University of Cambridge, in his 2008 book *100 Essential Things You Didn't Know You Didn't Know*. Barrow explains how to hedge your cash around different bookies to ensure that whatever the outcome of the race, you make a profit.

Although each bookie will stack their own odds in their favour, thus ensuring that no punter can place bets on all the runners in a race and guarantee a profit, that doesn't mean their odds will necessarily agree with those of a different bookie, says Barrow. And this is where gamblers can seize their chance.

Let's say, for example, you want to bet on one of the highlights of the British sporting calendar, the annual university boat race between old rivals Oxford and Cambridge. One bookie is offering 3 to 1 on Cambridge to ›

How to beat the bookies

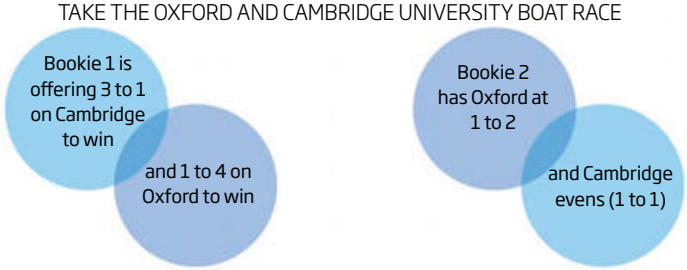
SUPPOSE THERE IS A RACE WITH *N* RUNNERS

You can always make a profit if Q is less than 1,

where $Q = \frac{1}{(a_1+1)} + \frac{1}{(a_2+1)} + \ldots + \frac{1}{(a_N+1)}$; a_1 is the odds on runner 1, a_2 is the odds on runner 2, etc

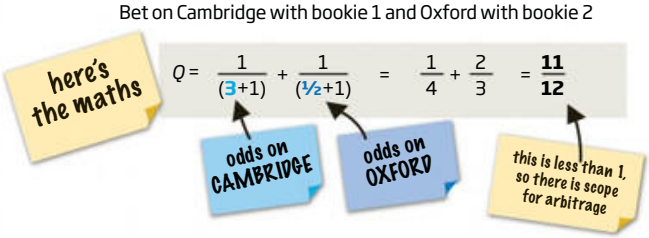
If $Q < 1$ there is an arbitrage opportunity. You can take advantage of it by gambling $\left(\frac{\frac{1}{a_1+1}}{Q}\right)$ of your money on runner 1, $\left(\frac{\frac{1}{a_2+1}}{Q}\right)$ on runner 2, and so on

AS A SIMPLE EXAMPLE,
TAKE THE OXFORD AND CAMBRIDGE UNIVERSITY BOAT RACE

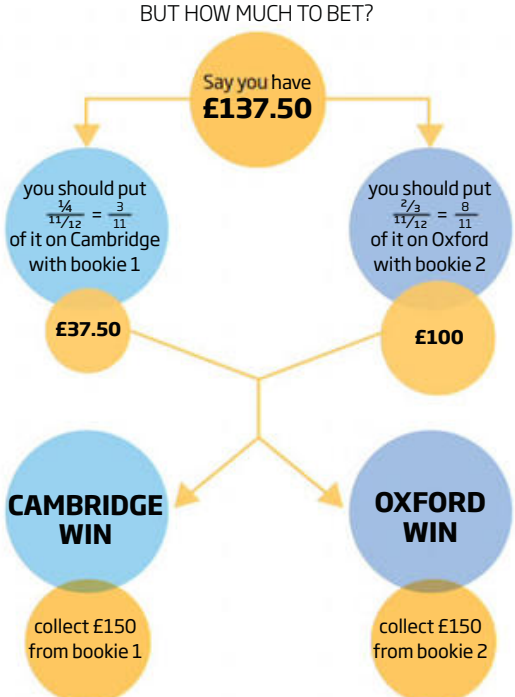


YOU CAN GUARANTEE PROFIT

Bet on Cambridge with bookie 1 and Oxford with bookie 2



BUT HOW MUCH TO BET?



(£112.50 winnings plus the £37.50 stake) You walk away with a **profit of £12.50**

(£50 winnings plus the £100 stake) You walk away with a **profit of £12.50**

win and 1 to 4 on Oxford. But a second bookie disagrees and has Cambridge evens (1 to 1) and Oxford at 1 to 2.

Each bookie has looked after his own back, ensuring that it is impossible for you to bet on both Oxford and Cambridge with him and make a profit regardless of the result. However, if you spread your bets between the two bookies, it is possible to guarantee success (see diagram, left). Having done the calculations, you place £37.50 on Cambridge with bookie 1 and £100 on Oxford with bookie 2. Whatever the result you make a profit of £12.50.

Simple enough in theory, but is it a realistic situation? Yes, says Barrow. "It's very possible. Bookies don't always agree with each other."

Guaranteeing a win this way is known as "arbitrage", but opportunities to do it are rare and fleeting. "You are more likely to be able to place this kind of bet when there are the fewest possible runners in a race, therefore it is easier to do it at the dogs, where there are six in each race, than at the horses where there are many more," says Barrow.

Even so, the mathematics is relatively simple, so I decided to try it out online. The beauty of online betting is that you can easily find a range of bookies all offering slightly

If the casino suspects you of card counting, you'll be asked to leave the blackjack table

different odds on the same race. "There are certainly opportunities on a daily basis," says Tony Calvin of online bookie Betfair. "It's not necessarily risk-free because you might not be able to get the bet you want exactly when you need it, but there are certainly people who make a living out of arbitrage."

After persuading a few friends to help me try an online bet, we followed a race, each keeping track of a horse and the odds offered by various online bookies. Keeping track of the odds to spot arbitrage opportunities was hard enough. Working out what to bet and when was, unsurprisingly, even harder. Arbitrage is not for the uninitiated.

"Arbitrage is not risk free because you might not be able to get the bet you want exactly when you need it. But there are certainly people who make a living out of it"

Cut your losses

However, it's still quite addictive, especially when you get tantalisingly close to finding a winning combination. And that's the problem with gambling – even when you have got mathematics on your side, it's all too easy to lose sight of what you could lose. Fortunately, that's the final thing that probability can help you with: knowing when to stop.

Everything in life is a bit of a gamble. You could spend months turning down job offers because the next one might be better, or keep laying bets on the roulette table just in case you win. Knowing when to stop can be as much of an asset as knowing how to win. Once again, mathematics can help.

If you have trouble knowing when to quit, try getting your head around "diminishing returns" – the optimal stopping tool. The best way to demonstrate diminishing returns is the so-called marriage problem. Suppose you are told you must marry, and that you must choose your spouse out of 100 applicants. You may interview each applicant once. After each interview you must decide whether to marry that person. If you decline, you lose the opportunity forever. If you work your way through 99 applicants without choosing one, you must marry the 100th. You may think you have 1 in 100 chance of marrying your ideal partner, but the truth is that you can do a lot better than that.

If you interview half the potential partners then stop at the next best one – that is, the first one better than the best person you've already interviewed – you will marry the very best candidate about 25 per cent of the time. Once again, probability explains why. A quarter of the time, the second best partner will be in the first 50 people and the very best in the second. So 25 per cent of the time, the rule "stop at the next best one" will see you marrying the best candidate. Much of the rest of the time, you will end up marrying the 100th person, who has a 1 in 100 chance of being the worst, but hey, this is probability, not certainty.

You can do even better than 25 per cent, however. John Gilbert and Frederick Mosteller of Harvard University proved that you could raise your odds to 37 per cent by interviewing 37 people then stopping at the next best. The number 37 comes from dividing 100 by e, the base of the natural logarithms, which is roughly equal to 2.72. Gilbert and Mosteller's law works no matter how many candidates there are – you simply divide the number of options by e. So, for example, suppose you find 50 companies that offer car insurance but you have no idea whether the next quote will be better or worse than the previous one. Should you get a quote from all 50? No, phone up 18 (50 ÷ 2.72) and go with the next quote that beats the first 18.

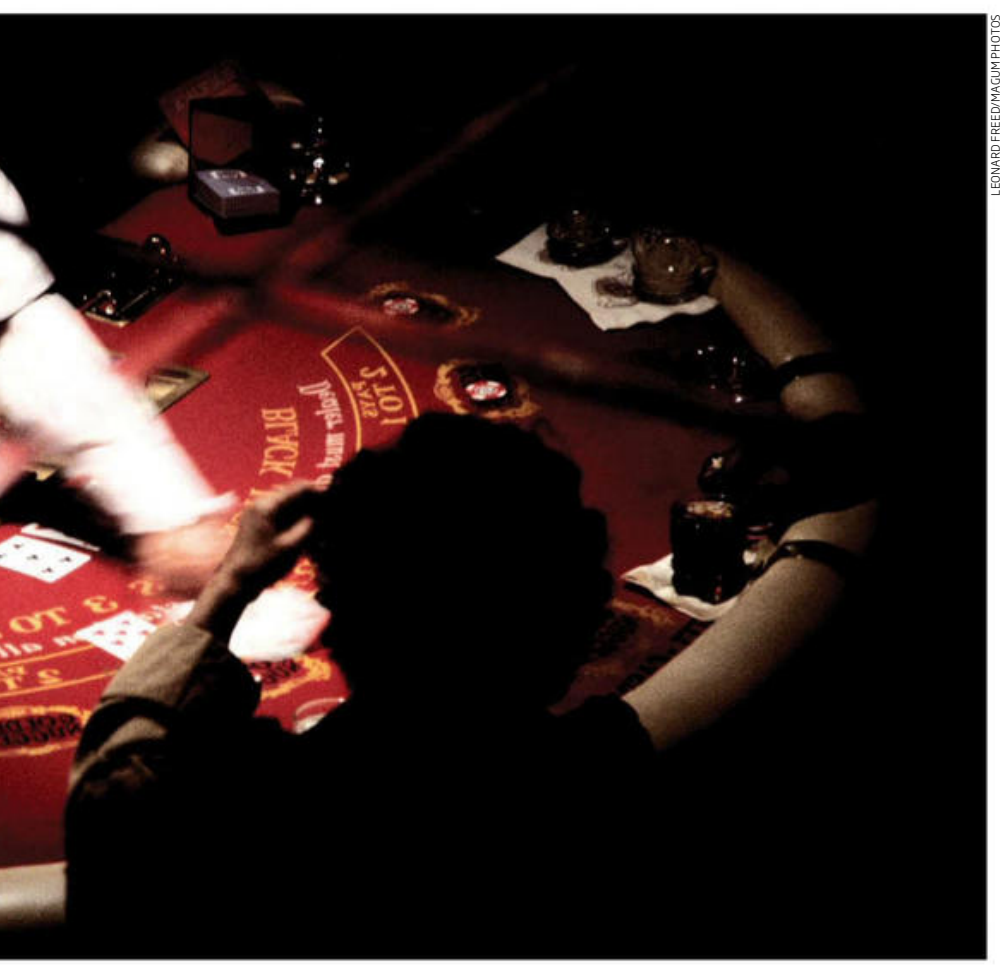
This can also help you decide the optimal time to stop gambling. Say you fancy placing some bets at the bookies. Before you start, decide on the maximum number of bets you will make – 20, for example. To maximise your chance of walking away at the right time, make seven bets then stop at the next one that wins you more than the previous biggest win.

Sticking to this rule is psychologically difficult, however. According to psychologist JoNell Strough at West Virginia University in Morgantown, the more you invest, the more likely it is that you will make an unwise decision further down the line.

This is called the sunk-cost fallacy, and it reflects our tendency to keep investing resources in a situation once we have started, even if it's going down the pan. It's why you are more likely to waste time watching a bad movie if you paid to see it.

So if you must have a gamble, use a little mathematics to give you a head start, or at least to tell you when to throw in the towel. Personally I think I'll retire. Overall I'm £11.50 up – a small win at the casino offset by losing £1 on my lottery ticket. It was a lot of effort for little more than pocket change.

Maybe I should have just put it all on red. ■



LEONARD FREED/MAGUM PHOTOS

GRAND DESIGNS

Symmetry isn't just for snowflakes and mirrors. Look deeper, says mathematician **Marcus du Sautoy**, and you find it rules the universe

OSLO, May 2008. King Harald of Norway presents mathematicians John Thompson and Jacques Tits with the Abel prize, one of the highest accolades in mathematics. There is a pleasing symmetry at the heart of the award. The winners are being honoured for ground-breaking work that led to the completion of a project started by Niels Abel, the 19th-century Norwegian mathematician after whom the prize is named. Appropriately enough, that project concerns mathematicians' attempts to answer the question: what is symmetry?

Most people's response is to point to the left-right reflectional symmetry of the human face. Or a flower, or a snowflake. But a snowflake has additional symmetries to that of a human face: as well as looking at its two halves, you can also turn a snowflake 60 degrees to match up its shape again. This begins to get at the essence of what symmetry is – a transformation or move that you can do to a structure that somehow makes it look like it did before you moved it. So how many other types of symmetry are there?

Remarkably, we now have a definitive answer. Thompson, of the University of Florida in Gainesville, and Tits, of the Collège de France in Paris, are responsible for ideas that have culminated in what is essentially a "periodic table" of symmetry. It has been as influential in the world of symmetry as the periodic table of elements has been to chemistry, allowing anyone exploring the complicated mathematical symmetries of an object to reduce it to something far simpler.

That matters because the symmetries of a structure often reveal secrets about how it behaves. For chemists, symmetry is key to classifying the possible crystals that can exist; in biology the mechanism of a virus owes much to its symmetrical shape; even the menagerie of fundamental particles revealed by physicists' super-colliders only make sense when you start to see them as facets of some strange, higher-dimensional symmetrical shape. And much of the technology we take for granted, such as mobile phones and the internet, depends on codes that exploit symmetry to preserve data as it is transmitted around the world.

Symmetry has fascinated civilisations since ancient times too, but it wasn't until the 19th century that we developed the language to understand its mathematics. This language allowed us to pull symmetry apart and discover its basic building blocks.

Just as molecules can be broken down into atoms like sodium and carbon, or numbers can be built out of the indivisible primes such as 3, 5 and 7, the mathematicians of Abel's generation discovered that symmetrical objects can be decomposed into indivisible symmetrical objects. Christened "simple groups", they are the atoms of symmetry.

Abel's contemporaries discovered that prime numbers are behind some of the first simple groups. Take a 15-sided polygon, for example. Its symmetries can be built from the symmetries of a pentagon and a triangle sitting inside the shape. To see how this works, imagine rotating the polygon by one-15th of a turn. Another way to do this is to first rotate the pentagon by two-fifths of a turn; then pull back in the opposite direction, rotating the triangle by one-third of a turn (see "Building blocks of symmetry", page 94). The reason this works is because $1/15 = 2/5 - 1/3$.

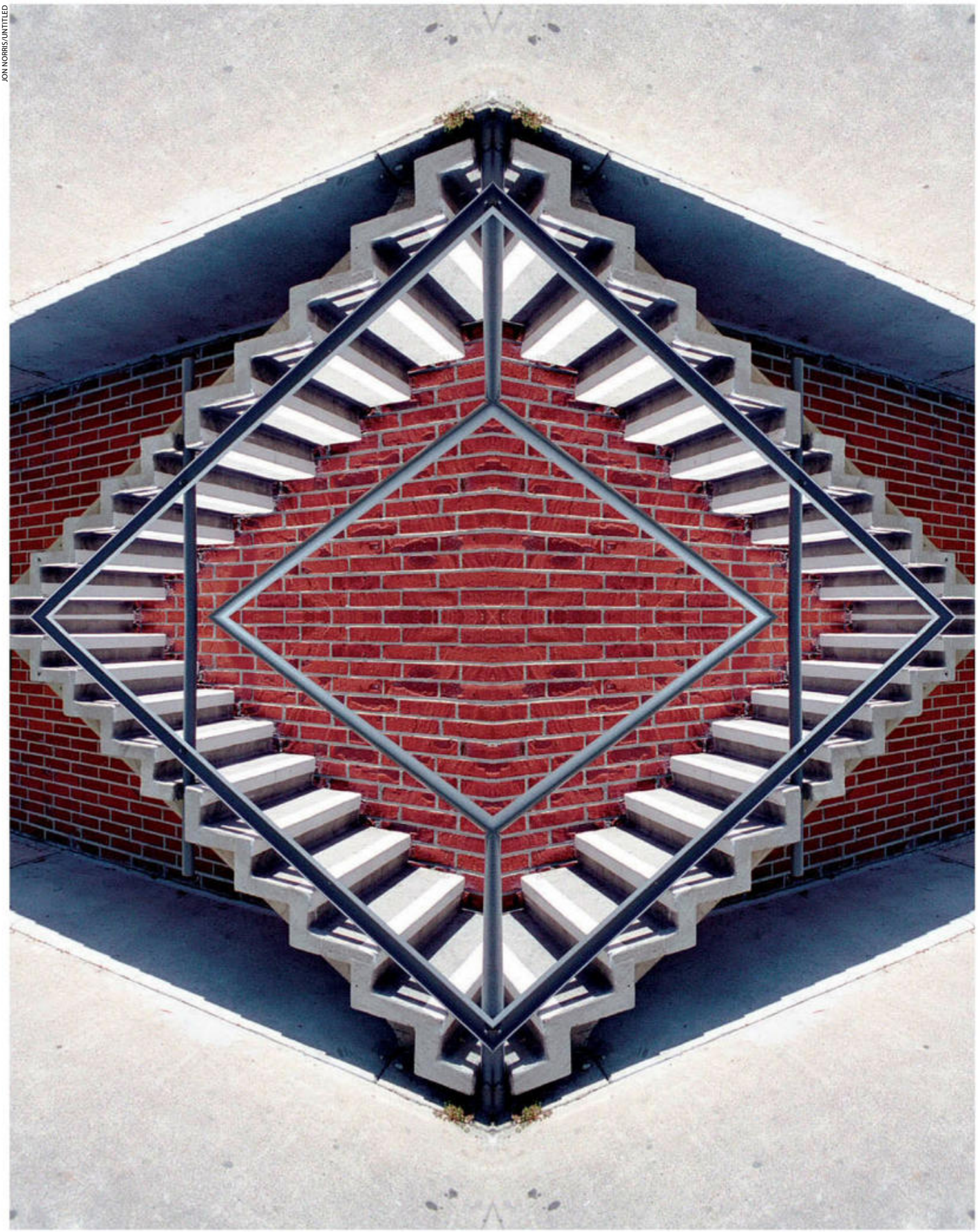
In fact, the symmetries of any flat, regular polygon can be broken down into the symmetries of the prime-sided shapes which fit inside the larger shape. For example, since $105 = 3 \times 5 \times 7$, the symmetries of a 105-sided figure are built from the symmetries of a triangle, a pentagon and a heptagon.

The Abel prize rewarded Thompson for a stunning theorem he proved with the late mathematician Walter Feit, showing that many more symmetrical objects in the mathematical world can be built out of prime-sided shapes. The beauty of their proof is that it applies to a plethora of shapes beyond simple 2D polygons. However complicated the shape, just knowing that it has an odd number of symmetries is enough to show that it could be pulled apart.

Thompson and Feit's theorem was impressive not just because it was a massive stepping stone to understanding the world of symmetry but also because the proof itself was massive. Called the odd order theorem, the 1963 paper describing it ran to 255 pages. At the time, it was possibly the longest proof that had ever been published.

So prime-sided polygons are the first indivisible symmetrical objects in the mathematician's periodic table of symmetry, but they are not the only ones. More exotic shapes were uncovered by 19th-century mathematicians when they tried to crack one of the other big problems of the day.

They knew of formulae that allowed ›



JON NORRIS/UNTITLED

them to work out solutions to equations involving x^2, x^3 or x^4. But they could not find a formula for solving "quintic" equations that include x raised to the fifth power, such as $x^5 + 6x + 3 = 0$.

Abel discovered the reason why a formula was so hard to pin down was that there wasn't one. A young French mathematician called Évariste Galois then found a way to push Abel's ideas a step further and give some basis as to why this could be. His revolutionary realisation was that behind every equation there is a symmetrical object.

The first hint of symmetry at work is evident in simple equations such as $x^2 = 4$. Its two solutions, $x = 2$ and $x = -2$, are in some ways mirror images of each other. A cubic equation has three solutions, and these are connected by the symmetries of a triangle. Once you get to quartic equations like $x^4 - 5x^3 - 2x^2 - 3x - 1 = 0$, the four solutions are connected by the symmetries of a tetrahedron, the 3D shape made from piecing together four equilateral triangles.

What Galois discovered is that if the symmetries of the object behind the equation can be broken down into prime-sided shapes, then a formula for finding the equation's solutions does exist. This unexpected connection with solving equations was the first indication that symmetry could be the key to unlocking many questions that did not at first sight seem to have anything to do with the concept.

When Galois began to examine quintic equations, he discovered that the symmetrical object at their heart is the dodecahedron, the 3D shape built from 12 pentagons. And the reason there is no formula for solving quintic equations is because the rotational symmetries of the dodecahedron cannot be broken up into prime-sided shapes.

There are 60 different ways to spin a dodecahedron so that all the pentagonal faces line up as they did before it was spun: in the language of mathematics, a dodecahedron has 60 different rotational symmetries. Even though 60 is a highly divisible number, Galois proved that the group of 60 rotations of the dodecahedron are as indivisible as if the shape were prime-sided.

Try to break the group of symmetries of the dodecahedron apart by using the rotations of one of its pentagonal faces and the result makes no sense. You can turn the object by one-fifth of a turn about a face, but there are no other shapes whose symmetries can be combined with those of the pentagon to build the symmetries of the dodecahedron.

Having discovered that the dodecahedron has so much in common with prime-sided shapes such as triangles and heptagons, the hunt was then on to find all indivisible shapes.

Mathematicians began to move away from physical objects and turned instead to more abstract structures. Remarkably, they found that shuffling a deck of cards behaves very much like the rotations or reflections of a physical shape. To understand how this works, start off by imagining a tetrahedron. Its rotational symmetry means that there are 12 ways of placing it on its triangular base so that it looks the same, as well as 12 reflectional symmetries. Now imagine sticking a jack, queen, king and ace of spades on the faces. When you rotate the tetrahedron, the movements are equivalent to shuffling the deck of cards. In fact, the symmetries of a tetrahedron can be modelled by the shuffles of a deck of four cards, which can take any one of 24 possible combinations. Similarly, the 60 symmetries of a dodecahedron are intimately related to the shuffles of five cards. What is powerful about this approach is that even though the number of three-dimensional shapes is limited, we can keep on exploring symmetries by adding cards to the deck.

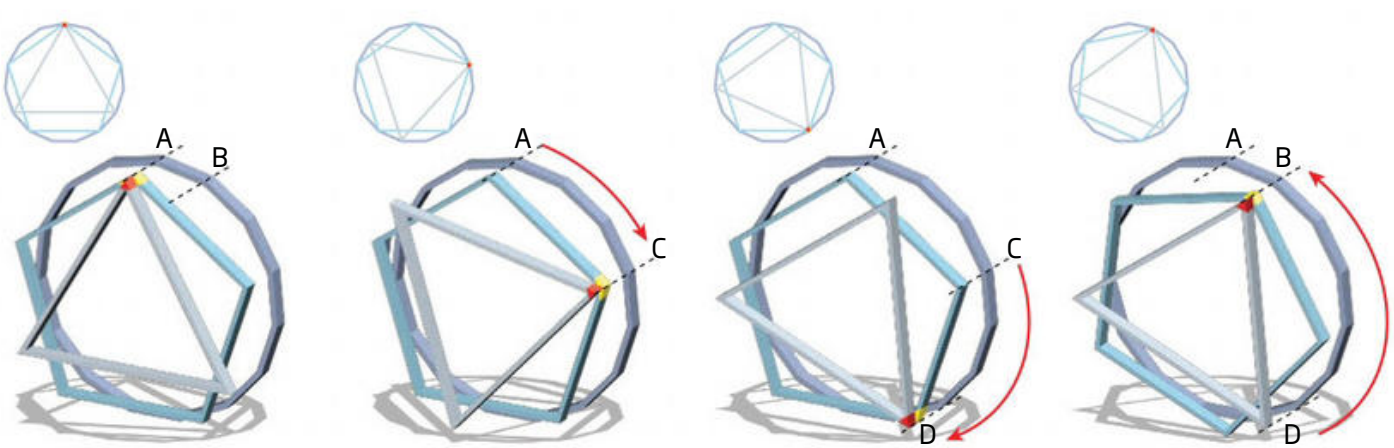
By changing their perspective and moving from 3D shapes to packs of cards, mathematicians discovered that the dodecahedron is not an isolated shape but the beginning of a new infinite family of indivisible symmetrical structures to add to the periodic table of symmetry alongside the prime-sided shapes.

Monster of the deep

So far, so good. But even greater rewards lie in navigating beyond the third dimension

Building blocks of symmetry

The symmetries of a 15-sided polygon can be built by combining the rotational symmetries of a triangle and a pentagon. To make $^1/_{15}$th of a turn of the polygon, from A to B, turn the pentagon by two-fifths (A to C, then C to D) taking the triangle with it. Then rotate the triangle by one-third in the opposite direction (D to B) taking the pentagon with it



and into hyperspace. The symmetries of these higher-dimensional shapes are the key to unlocking the behaviour of fundamental particles and building the standard model of particle physics. Symmetries are also behind many of the fundamental conservation laws in physics, as the German mathematician Emmy Noether discovered in 1915 (see "The hidden law", page 97).

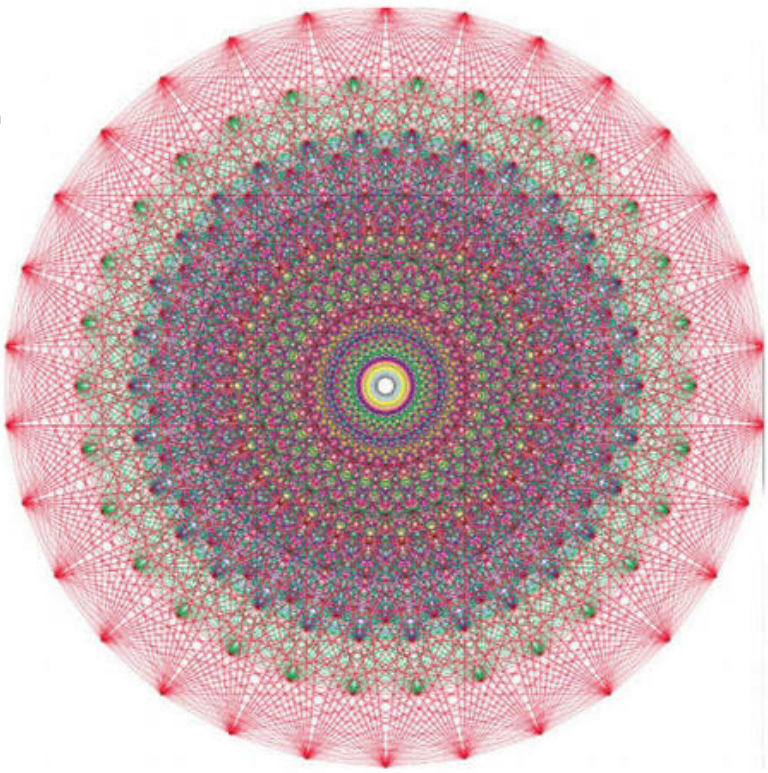
The reason mathematicians can manipulate objects in hyperspace is because of Descartes's dictionary, which turns geometry into numbers. Just as every position on the globe can be specified by two map coordinates, we can translate shapes into numbers. A square, for example, can be described by the coordinates of its corners: (0,0), (1,0), (0,1) and (1,1). Add an extra coordinate and you add a dimension, so you can then specify the eight corners of a cube as (0,0,0), (0,0,1) and so on.

So what about a four-dimensional cube? Although the pictures run out, the numbers don't, and this allows us to explore the geometry and symmetry of this shape. A four-dimensional cube is known as a tesseract and has 16 corners, 32 edges, 24 square faces and is constructed out of eight three-dimensional cubes. Its symmetries turn out to be related to another family of indivisible symmetries called simple groups of Lie type, after another Norwegian mathematician, Sophus Lie.

The symmetries of "hypercubes" are behind one of 16 new families of Lie groups. And it is unlocking the secrets of these groups for which the Belgian-born mathematician Tits was recognised with the award of the Abel prize. Tits constructed geometrical settings in higher dimensions that help explain the symmetries of these families.

There are more indivisible symmetries to add to the periodic table, but the other groups aren't as well behaved as the Lie groups, shuffles or prime-sided shapes. At the end of the 19th century, a French mathematician called Emile Mathieu had discovered five indivisible symmetries that didn't seem to fit into any of these patterns, nor did they create a family of their own. They just seemed to be sitting there like orphans. Were these five the only exceptional groups of symmetries, or what mathematicians called sporadic groups?

In 1965, Thompson received a letter from the Croatian mathematician Zvonimir Janko, who claimed to have discovered a sixth sporadic group. At first Thompson was quite dismissive of the claim, but as he analysed Janko's proposal he realised the Croatian could be onto something.



The symmetry group E8 reveals a deep relationship between the universe's forces and particles

"The periodic table of symmetry is as influential in mathematics as its namesake in chemistry"

Janko's discovery turned out to be the beginning of a crazy period in the story of symmetry when mathematicians discovered a whole range of strange indivisible sporadic groups of symmetry that didn't seem to fit any of the patterns determined by previous generations. Many of the discoveries depended on using a formula developed by Thompson to predict how many symmetries such a sporadic group might have.

Often the birth of these sporadic groups mirrored the discovery of fundamental particles in physics. By exploiting the symmetries underlying the standard model of particle physics, theorists predicted the existence of particles such as the charm quark several years before experiments found evidence for them. Similarly, mathematicians used Thompson's formula to predict objects before they were actually constructed.

Thompson and Tits are among those who have their names attached to some of these sporadic groups. The culmination of this period of exploration was the prediction by German mathematician Bernd Fischer of an object that can only be seen from 196,883-dimensional space and has more symmetries than there are atoms in the sun. Robert Griess, a mathematician at the University of Michigan in Ann Arbor, eventually constructed the object in 1980.

Called simply "the monster", it is the largest of the sporadic groups. Far from being some anomalous freak with no relation to reality, we are beginning to realise that the symmetries of the monster might actually underpin some of the deepest ideas of string theory – currently our best hope of uniting relativity and quantum physics.

We are finally coming to the realisation that the monster was the last: there are no more indivisible symmetries to add to the periodic table of symmetry. In what many regard as one of the greatest achievements of mathematics, we now have a complete list of the building blocks of symmetry. It is the power of mathematical proof that we can be so sure that the list is complete, but it is thanks to the work of mathematicians like Thompson and Tits that we are able to produce such a definitive answer. It's now up to the next generation to explore what symmetrical objects we can build from these atoms of symmetry. ■

AIMATH

STUART FRANKLIN / MAGNUM PHOTOS

The hidden law

The deep connection between mathematical symmetry and physical laws was the penetrating insight of an overlooked genius a century ago, says **Dave Goldberg**

WE PHYSICISTS have a habit of depicting our discipline as "beautiful" or "elegant", where an outsider might be forgiven for seeing no more than an endless morass of equations. In an ideal world, those equations would be unnecessary; the ultimate goal of physics – and science generally – is to describe the world as simply as possible.

In 1915 Albert Einstein caught the attention of the mathematical world with the presentation of his general theory of relativity. But that same year, the excitement surrounding relativity spawned another seminal piece of work. Even among physicists, though, it is not nearly as famous as it should be. Perhaps that is down to the complexity of its mathematics, but perhaps the author's sex and sadly short life played their parts too.

Yet there is no doubt that Amalie "Emmy" Noether transformed how we think about the universe. Despite the hairy mathematics, her great first theorem can be described conceptually in just a short sentence: *Symmetries give rise to conservation laws.*

This simplicity masks a penetrating insight. It provided a unifying perspective on the physics known at the time – and laid the groundwork for nearly every major ›

IAN BERRY/MAGNUM PHOTOS

For physicists, the appeal of symmetry goes beyond the purely aesthetic

fundamental discovery since.

Emmy Noether is a story unto herself. Despite wide recognition of her obvious brilliance, she was confounded by the prejudices of German academic tradition at the turn of the 20th century. Born into a prominent mathematical family in 1882 – her father, Max, was a professor at the University of Erlangen in the north of Bavaria – she was at first forbidden from enrolling at the university because of her gender.

Even though Noether was eventually able to gain both an undergraduate degree and a PhD, still no university would hire her for their faculty. Over the next decade, she became one of the world's experts in the mathematics of symmetry – but without appointment, pay or formal title.

Symmetry may seem like a trifling subject at first blush. The mathematician Hermann Weyl, a contemporary of Noether's who was greatly influenced by her work, once described a very simple way of thinking about the concept: "A thing is symmetrical if there is something you can do to it so that after you have finished doing it, it looks the same as before," he wrote. A circle, for instance, can be rotated by any angle and looks the same.

The idea that symmetries lie at the heart of physical laws is old. Aristotle and his contemporaries argued that the stars were pasted on celestial spheres, and that the globes moved in circular orbits. They were wrong, as it happens. As Johannes Kepler discovered through meticulous observation in the early 17th century, planets wander closer and further from the sun, in the geometric form of an ellipse. They travel faster when closer in, and slower when further out. An imaginary line connecting planets to the sun traces out equal areas in equal times: what we now know as conservation of angular momentum.

Beyond relativity

It wasn't until later that century that Newton explained why this happens, with his universal law of gravitation. The source of this behaviour was indeed a symmetry – the symmetry of the invisible hand of gravity, which acts equally in all directions from a massive body such as the sun.

General relativity, Einstein's much refined theory of gravity, was founded on a symmetry too, one known as the equivalence principle. This states that there is no practical difference between a body experiencing acceleration because of gravity and one experiencing an equivalent acceleration from a different source, such as the thrust of a rocket or the spin of a centrifuge. From the equivalence principle, Einstein developed his theory that yields everything from curved space-time and an expanding universe to black holes and the prediction, unconfirmed until 2015, of gravitational waves rippling through space.

Einstein's work revolutionised our view of the universe, but also spurred a great deal of interest in the role of symmetries in physical laws. Recognising Noether as an expert, in 1915 the eminent mathematicians David Hilbert and Felix Klein invited her to Göttingen, then the centre of the mathematical world – an offer, alas, which still didn't extend to any financial remuneration. Hilbert did argue forcefully for an official appointment, but Noether wasn't given even an honorary "extraordinary" professorship until 1922. In the interim, she was merely allowed to serve as a guest lecturer, unpaid, under Hilbert's name.

Weyl, also at Göttingen in the 1920s, by

contrast quickly achieved a prominent professorship, despite being Noether's junior. "I was ashamed to occupy such a preferred position beside her whom I knew to be my superior as a mathematician in many respects," he later remarked.

The indignities of Noether's circumstances did not deter her work. Almost immediately on arrival, Noether developed her eponymous theorem. It formalised the idea, intrinsic but unstated in the examples of the two theories of gravity, that symmetries provide an express route to the heart of nature's workings.

For another example, consider a puck placed on a very smooth, very large frozen lake. Wherever the puck slides, the lake is the same. Noether's theorem provides a general way of turning that statement of symmetry into a conservation law.

Conservation laws are the bread and butter of physics. They are mathematical shortcuts that allow us to compute physical quantities once and then never again. Whatever you start with, that's what you'll end up with. That is incredibly useful: think how much trickier it would be to manage your time if the number of hours in the day changed constantly and were not conserved at 24; it's bad enough twice in the year when the clocks go forward or back.

Most of the great laws of physics include some statement of conservation, implicitly or explicitly. Newton's first law of motion crudely states that "objects in motion stay in motion, and objects at rest stay at rest". That is nothing more than conservation of momentum, a consequence of the sort of spatial symmetry

"THE INDIGNITY OF NOETHER'S CIRCUMSTANCES DID NOT DETER HER WORK"

that governs the physics on top of our idealised frozen lake. Send a puck across the ice and, discounting friction, it will continue indefinitely. But the conservation law only holds as far as the symmetry does. A hole in the ice will disturb the symmetry, causing the puck to sink to the bottom of the lake and come to rest – violating Newton's first law.

It's not always obvious what is conserved and what isn't. For a long time, it was assumed that mass couldn't be created or destroyed, but Einstein's famous relation $E=mc^2$ said otherwise. Matter can be created, if not out of thin air, then out of pure energy. In fact, although you are made of molecules that are made of protons and neutrons, those protons and neutrons are made of quarks. Quarks, as it happens, are so light that they make up only about 1 to 2 per cent of your body mass. The rest comes from the incredible energies with which these quarks interact.

Although matter can be created from energy, energy itself in all its myriad forms adds up to a constant and permanent total. Before Noether, energy was simply assumed to be conserved, an assumption so basic that it became known in the 19th century as the first law of thermodynamics. But do the mathematics associated with Noether's theorem and it becomes plain that energy is conserved because of an even more basic symmetry: specifically, that the laws of physics aren't changing with time. If they did, energy wouldn't be conserved.

What Noether's theorem adds up to is a practical prescription for making progress in physics: identify a symmetry in the world's workings, and the associated conservation law will allow you to start meaningful calculation.

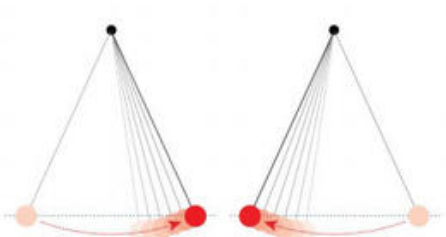
But it is also, in a sense, a statement about how the universe *should* be structured. When we look at the universe on the human scale, or even at the level of our solar system, space seems very different from a smooth lake: there are planet-sized bumps and wiggles. But take a broader picture – on the scale of hundreds of millions of light-years – and the universe appears much smoother. The assumption is that on the very largest scales, ›

Magic recipe

Symmetries exist everywhere in nature. Emmy Noether's theorem of 1915 provides a way to translate them into laws useful for calculations

SYMMETRY: TRANSLATION IN TIME
The basic laws of physics do not vary over time

CONSEQUENCE: Energy is conserved
However many times a pendulum swings, with no air resistance it will always reach the same height

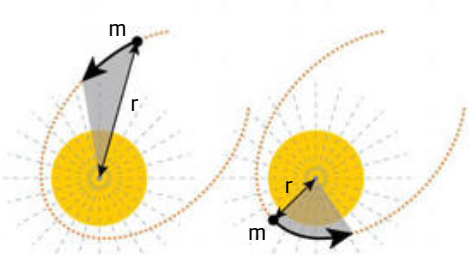


SYMMETRY: TRANSLATION IN SPACE
The laws of physics don't change when you move from one place to another

CONSEQUENCE: Momentum is conserved
A rocket flying through free space continues flying at the same speed, if no other forces act on it

SYMMETRY: ROTATION IN SPACE
Forces such as gravity emanate equally in all directions

CONSEQUENCE: Angular momentum is conserved
Comets speed up nearer the sun. The area between their path and the sun is always the same in a set time



the universe is more or less the same.

As we lack the ability to travel billions of light years to beyond the observational horizon of our most powerful telescopes, this really is just an assumption, and it goes by the name of the cosmological principle. It tells us that what we call "down" on Earth is nothing more than a consequence of the relative position of us and the rock we're standing on. The universe has no up or down, nor a centre for that matter. Its laws don't seem to be in any way related to where we measure them, how our measuring devices are pointed, or even when we decide to make the measurements. Through Noether's theorem, symmetries of space and time yield conservation of energy, momentum and angular momentum everywhere, all the time (see "Magic recipe", page 99).

But there's much more. Symmetries in space and time might be obvious to the naked eye, yet Noether's theorem's true strength comes from "internal symmetries". To the uninitiated, the standard model of particle physics is nothing more than a list of fundamental forces and particles. But it is a model of internal symmetries writ large, and it was built on Noether's theorem.

The most familiar force it deals with is electromagnetism, which describes the current running through our power cords, the behaviour of compasses and the shock of lightning. James Clerk Maxwell is generally credited for writing down a theory that unified electricity and magnetism into one working model in the 1860s. One of its assumptions is that electric charge is neither created nor destroyed, an idea that goes back even further to Benjamin Franklin in the 1740s.

"WE ASSUME THE UNIVERSE HAS NO UP OR DOWN, NOR A CENTRE FOR THAT MATTER"

Noether's theorem shows that charge conservation, too, arises from a symmetry. Fundamental particles have a property called spin, and just as position doesn't matter on a frozen lake, what's known as the spin's phase doesn't change physical calculations. "Turn" every electron in the universe an extra degree, and neither energy nor anything else changes. What pops out, according to Noether's mathematics, is charge conservation.

Weyl took this idea of phase symmetry a step further and supposed that every electron might be twisted by a different amount and still remain the same. Assume this and, almost by magic, all four of Maxwell's equations emerge.

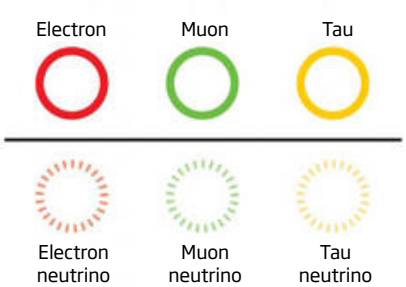
As the standard model has developed, so the symmetries of interest have become more subtle – but Noether's theorem has been the gift that keeps on giving. It is hard to conceive, for example, that electrons, the particles that run through wires to power electronics, and neutrinos, which fly through us by the trillions every second without leaving a mark, are in some sense the same particle.

Neutrinos primarily interact through the weak force, which controls nuclear fusion in the sun. But the weak force is indifferent to

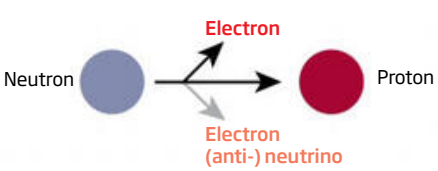
Hidden principles

The workings of the standard model of particle physics – and perhaps physics theories beyond it – are determined by some subtle symmetries

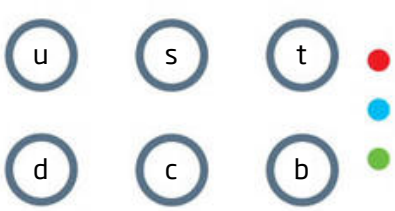
SYMMETRY: the electron and similar charged "leptons" have a neutral neutrino equivalent



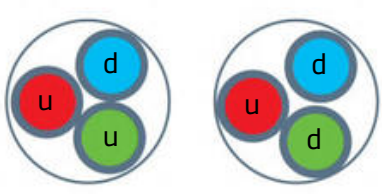
To keep the balance when electrons are emitted in processes such as beta decay, a neutrino is also produced



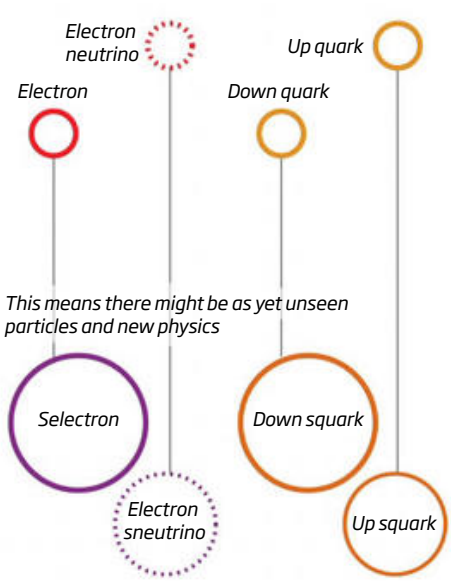
SYMMETRY: all of the six varieties of quark come in the same "colours"



Particles containing quarks, such as protons and neutrons, are made of "colour-neutral" combinations. Switch two colours and nothing changes



SYMMETRY: *according to the unproven theory of supersymmetry, every lepton and quark should have a heavier "sparticle" equivalent*



This means there might be as yet unseen particles and new physics

whether a particle is an electron or a neutrino: switch them round, and weak interactions will be the same. This symmetry produces conservation of a quantity called weak isospin that, like electric charge, can be used to label particles and predict how they will behave (see "Hidden principles", below).

In the 1960s, researchers found that electromagnetism and the weak force could in fact be generated by a single symmetry, in what became known as the electroweak theory, a keystone of the standard model. "Breaking" that symmetry into two separate pieces produced a bunch of new interactions, along with the prediction of a new particle – what we now know as the Higgs boson. We waited a half-century for the confirmation of this prediction, which stemmed directly from the sort of considerations Noether's theorem introduced into physics. It came, eventually, with the discovery of the Higgs at CERN's Large Hadron Collider in 2012.

The other pillar of the standard model is the strong interaction, which holds individual protons and neutrons together. The quarks that make up these particles are labelled with one of three "colours": red, green and blue. Shift all the colours by one, and all strong interactions will remain exactly the same.

Colour symmetry leads – in what might at first seem to be a tautology – to conservation of colour. Since that idea was first introduced, work on the nature of the strong force has found that all particles in nature exist in states without colour – "white", effectively. Protons and neutrons are examples of particles called baryons that consist of three quarks, one red, one blue, one green. The universe as a whole seems to be colourless, just as it is electrically neutral, and the symmetry of the strong force is what makes particles like protons and neutrons possible in the first place.

THE GRANGER COLLECTION/TOPFOTO

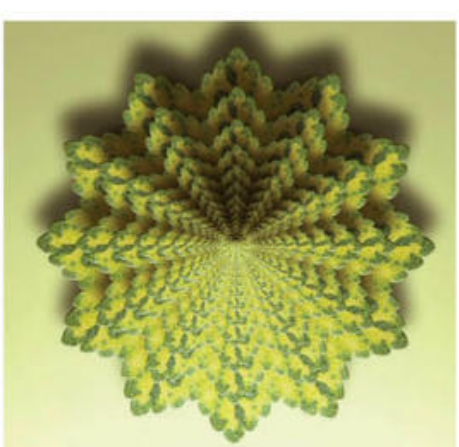
Emmy Noether remains largely unknown, despite her seminal work

The thrill of the chase

Physics is now at the point where new theories are built on the assumption of a fundamental symmetry, and an informed guess about what that symmetry might be. Unification is a holy grail of physics: the drive to develop theories that can describe everything in just a few, albeit possibly outstandingly difficult, equations. What sort of symmetry might unify the electroweak and strong forces we do not yet know, but the search for such a "grand unified theory" is an active area of physical endeavour. A good grand unified theory might predict where all of the protons and neutrons in the universe come from.

The total number of these baryons seems to be conserved too. Experimentally, we've tried to see if protons, the lightest of the baryons, can decay into anything. If we ever observe this we will have some idea as to whether baryon number is really conserved, a key clue to a grand unified theory.

Of particular interest as we look beyond the standard model is supersymmetry, a model at the heart of many fledgling grand unified theories. Supersymmetry is based on unifying the two major groups of fundamental particles: fermions (the particles that make up matter such as electrons and quarks), and bosons (including the photon, the Higgs and other particles governing forces). It supposes that ultimately every fermion has a partner boson and vice-versa:



GALLERYSTOCK

hypothetical exotics such as "selectrons" and "higgsinos". At high enough energies, the supposition is that an electron and a selectron behave the same way, just as neutrinos and electrons behave identically under the weak force.

Supersymmetry neatly solves many problems of the standard model, as well as providing a motivation for why particles have the masses that they do. In principle, that is. The Large Hadron Collider is hard at work looking for signatures of supersymmetry, but the lack of any success so far suggests we're barking up the wrong tree.

Even further away is the goal of folding gravity, that original object of symmetric study, and the forces covered by the standard model into a "theory of everything". Indeed, physics is still far away from a final resolution. But in the thrill of the chase for better answers, it is studying symmetries that will guide us along the way – and it is Noether's theorem that will magic useful physical insights from that.

Compared with this stellar legacy, the rest of Noether's biography is kind of a downer. She left Germany to escape the Nazis in 1933 and came to Bryn Mawr College in Pennsylvania, dying of complications from cancer surgery two years later. As Einstein wrote after her death, "Fräulein Noether was the most significant creative mathematical genius thus far produced since the higher education of women began". Others might suggest the second part of that sentence is superfluous. ■

Universe by numbers

What is reality made of? For **Max Tegmark**, it's all in the maths

WHAT is the meaning of life, the universe and everything? In the sci-fi spoof *The Hitchhiker's Guide to the Galaxy*, the answer was found to be 42; the hardest part turned out to be finding the real question. Indeed, although our inquisitive ancestors undoubtedly asked such big questions, their search for a "theory of everything" evolved as their knowledge grew. As the ancient Greeks replaced myth-based explanations with mechanistic models of the solar system, their emphasis shifted from asking "why" to asking "how".

Since then, the scope of our questioning has dwindled in some areas and mushroomed in others. Some questions were abandoned as naive or misguided, such as explaining the sizes of planetary orbits from first principles, which was popular during the Renaissance. The same may happen to currently trendy pursuits like predicting the amount of dark energy in the cosmos, if it turns out that the amount in our neighbourhood is a historical accident. Yet our ability to answer other questions has surpassed earlier generations' wildest expectations: Newton would have been amazed to know that we would one day measure the age of our universe to an accuracy of 1 per cent, and comprehend the microworld well enough to make an iPhone.

Mathematics has played a striking role in these successes. The idea that our universe is in some sense mathematical goes back at least to the Pythagoreans of ancient Greece, and has spawned centuries of discussion among physicists and philosophers. In the 17th century, Galileo famously stated that the universe is a "grand book" written in the language of mathematics. More recently, the physics Nobel laureate Eugene Wigner argued in the 1960s that "the unreasonable effectiveness of mathematics in the natural sciences" demanded an explanation.

Here, I will push this idea to its extreme and argue that our universe is not just described by mathematics – it is mathematics. While this hypothesis might sound rather far-fetched, it makes startling predictions about the structure of the universe that could be testable by observations. It should also be useful in narrowing down what an ultimate theory of everything could look like.

The foundation of my argument is the assumption that there exists an external physical reality independent of us humans. This is not too controversial: I would guess that the majority of physicists favour this long-standing idea, though it is still debated. Metaphysical solipsists reject it flat out, and supporters of the so-called Copenhagen interpretation of quantum mechanics may reject it on the grounds that there is no reality without observation. Assuming an external reality exists, however, physics theories aim to describe how it works. Our most successful theories, such as general relativity and quantum mechanics, describe only parts of this reality: gravity, for instance, or the behaviour of subatomic particles. In contrast, the holy grail of theoretical physics is a theory of everything – a complete description of reality.

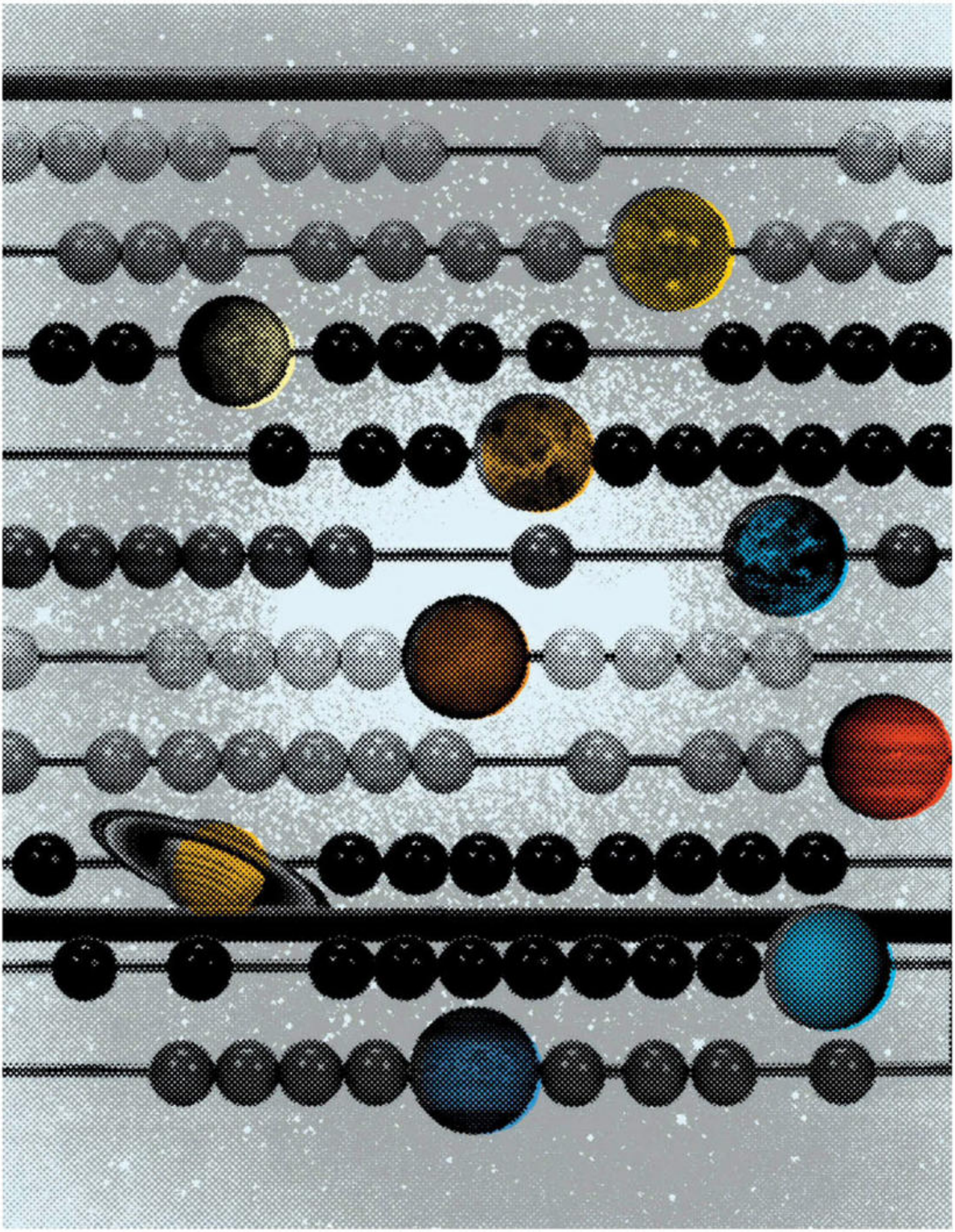
My personal quest for this theory begins with an extreme argument about what it is allowed to look like. If we assume that reality exists independently of humans, then for a description to be complete, it must also be well defined according to non-human entities – aliens or supercomputers, say – that lack any understanding of human concepts. Put differently, such a description must be expressible in a form that is devoid of human baggage like "particle", "observation" or other English words.

In contrast, all physics theories that I have been taught have two components: mathematical equations, and words that explain how the equations are connected to what we observe and intuitively understand. When we derive the consequences of a theory, we introduce concepts – protons, stars, molecules – because they are convenient. However, it is we humans who create these concepts. In principle, everything could be calculated without this baggage: a sufficiently powerful supercomputer could calculate how the state of the universe evolves over time without interpreting it in human terms.

All of this raises the question: is it possible to find a description of external reality that involves no baggage? If so, such a description of objects in this reality and the relations between them would have to be completely abstract, forcing any words or symbols to be mere labels with no preconceived meanings whatsoever. Instead, the only properties of these entities would be those embodied by the relations between them.

This is where mathematics comes in. To a ›

10^{57}
number of atoms in a typical star



AESTHETICAPPARATUS

modern logician, a mathematical structure is precisely this: a set of abstract entities with relations between them. Take the integers, or geometric objects like the dodecahedron, a favourite of the Pythagoreans (see diagram, below). This is in stark contrast to the way most of us first perceive mathematics – either as a sadistic form of punishment, or as a bag of tricks for manipulating numbers. Like physics, maths has evolved to ask broader questions.

Modern mathematics is the formal study of structures that can be defined in a purely abstract way. Think of mathematical symbols as mere labels without intrinsic meaning. It doesn't matter whether you write "two plus two equals four", "2 + 2 = 4" or "dos más dos igual a cuatro". The notation used to denote the entities and the relations is irrelevant; the only properties of integers are those embodied by the relations between them. That is, we don't invent mathematical structures – we discover them, and invent only the notation for describing them.

So here is the crux of my argument. If you believe in an external reality independent of humans, then you must also believe in what I call the mathematical universe hypothesis: that our physical reality is a mathematical structure. In other words, we all live in a gigantic mathematical object – one that is more elaborate than a dodecahedron, and probably also more complex than objects with intimidating names like Calabi-Yau manifolds, tensor bundles and Hilbert spaces, which appear in today's most advanced theories. Everything in our world is purely mathematical – including you.

If that is true, then the theory of everything must be purely abstract and mathematical. Although we do not yet know what the theory would look like, particle physics and cosmology have reached a point where all measurements ever made can be explained, at least in principle, with equations that fit on a few pages and involve merely 32 unexplained numerical constants. So the correct theory of everything could even turn out to be simple enough to describe with equations that fit on a T-shirt.

Before discussing whether the mathematical universe hypothesis is correct, however, there is a more urgent question: what does it actually mean? To understand this, it helps to distinguish between two ways of viewing our external reality. One is the outside overview of a physicist studying its mathematical structure, like a bird surveying a landscape from high above; the other is the inside view of an observer living in the world described by the structure, like a frog living in the landscape surveyed by the bird.

One issue in relating these two perspectives involves time. A mathematical structure is by definition an abstract, immutable entity existing outside of space and time. If the history of our universe were a movie, the structure would correspond not to a single frame but to the entire DVD. So from the bird's perspective, trajectories of objects moving in four-dimensional space-time resemble a tangle of spaghetti. Where the frog sees something moving with constant velocity, the bird sees a straight strand of uncooked spaghetti. Where the frog sees the moon orbit the Earth, the bird sees two intertwined spaghetti strands. To the frog, the world is described by Newton's laws of motion and gravitation. To the bird, the world is the geometry of the pasta.

A further subtlety in relating the two perspectives involves explaining how an observer could be purely mathematical. In this example, the frog itself must consist of a thick bundle of pasta whose structure corresponds to particles that store and process information in a way that gives rise to the familiar sensation of self-awareness.

Fine, so how do we test the mathematical universe hypothesis? For a start, it predicts that further mathematical regularities remain to be discovered in nature. Ever since Galileo promulgated the idea of a mathematical cosmos, there has been a steady progression of discoveries in that vein, including the standard model of particle physics, which captures striking mathematical order in the microcosm of elementary particles and the macrocosm of the early universe.

The hypothesis also makes a much more dramatic prediction: the existence of parallel universes. Many types of "multiverse" have been proposed over the years, and it is useful to classify them into a four-level hierarchy.

100 billion
number of stars in a typical galaxy

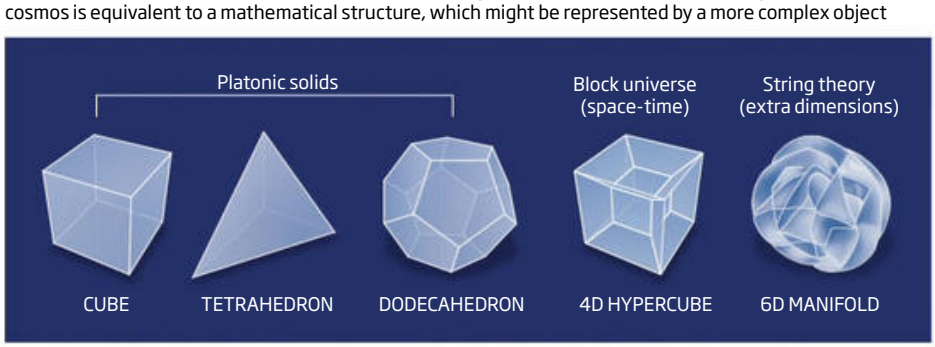
The first three levels correspond to non-communicating parallel worlds within the same mathematical structure: level I simply means distant regions from which light has not yet had time to reach us; level II covers regions that are forever unreachable because of the cosmological inflation of intervening space; and level III, often called "many worlds", involves non-communicating parts of the Hilbert space of quantum mechanics into which the universe can in a sense "split" during certain quantum events. Level IV refers to parallel worlds in distinct mathematical structures, which may have fundamentally different laws of physics.

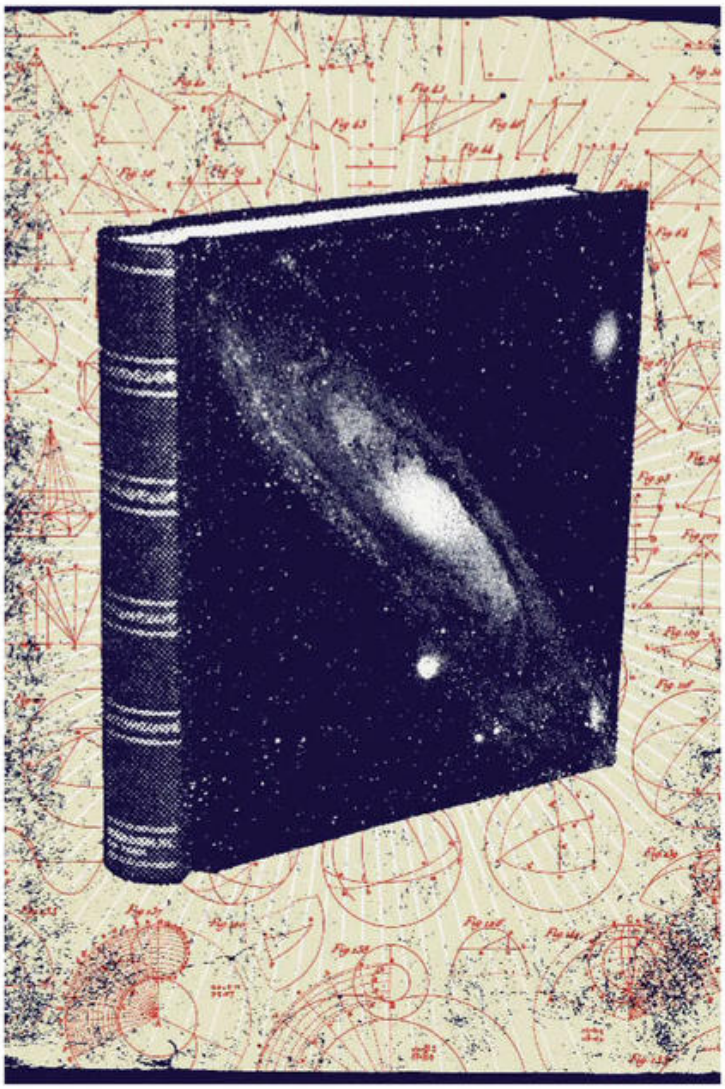
Today's best estimates suggest that we need a huge amount of information, perhaps 10^{100} bits, to fully describe our frog's view of the observable universe, down to the positions of every star and grain of sand. Most physicists hope for a theory of everything that is much simpler than this and can be specified in few enough bits to fit in a book, if not on a T-shirt.

The mathematical universe hypothesis implies that such a simple theory must predict a multiverse. Why? Because this theory is by definition a complete description of reality: if it lacks enough bits to completely specify our universe, then it must instead describe all possible combinations of stars, sand grains and such – so that the extra bits that describe our universe simply encode which universe we are in, like a multiversal phone number. Thus, describing a multiverse

Mathematical structures of the universe

Over the course of history, researchers have associated geometric shapes with properties of the universe. The ancient Greeks studied Platonic solids, while Einstein considered a 4D block universe and string theorists postulate extra dimensions of space. According to the "mathematical universe hypothesis", the cosmos is equivalent to a mathematical structure, which might be represented by a more complex object




AESTHETICA APPARATUS

80
number of stable chemical elements

can be simpler than describing one universe.

Pushed to its extreme, the hypothesis implies the level-IV multiverse. If there is a mathematical structure that is our universe, and its properties correspond to our physical laws, then each mathematical structure with different properties is its own universe with different laws. Indeed, the level-IV multiverse is compulsory, since these structures are not "created" and don't exist "somewhere" – they just exist. Stephen Hawking once asked, "What is it that breathes fire into the equations and makes a universe for them to describe?" For the mathematical cosmos, there is no fire-breathing required, since the point is not that a mathematical structure describes a universe, but that it is a universe.

The existence of the level-IV multiverse also answers a confounding question emphasised by the physicist John Wheeler: even if we found equations that describe our universe perfectly, then why these particular equations, not others? The answer is that the other equations govern parallel universes, and that our universe has these particular equations because they are statistically likely, given the distribution of mathematical structures that can support observers like us.

It is crucial to ask whether parallel universes are within the purview of science, or are merely speculation. They are not a theory in themselves, but rather a prediction made by certain theories. For a theory to be falsifiable, we need not be able to test all its predictions, merely at least one of them.

So here's a testable prediction: if we exist in many parallel universes, then we should expect to find ourselves in a typical one. Suppose we succeed in computing the probability distribution for some number, say the dark energy density or the number of dimensions of space, measured by a typical observer in a mathematical structure where this number has meaning. If we find that this distribution makes the value measured in our own universe highly atypical, it would rule out the multiverse, and hence the mathematical universe hypothesis.

Ultimately, why should we believe the mathematical universe hypothesis? Perhaps the most compelling objection is that it feels counter-intuitive and disturbing. I personally dismiss this as a failure to appreciate Darwinian evolution. Evolution endowed us with intuition only for those aspects of physics that had survival value for our distant ancestors, such as the parabolic trajectories of flying rocks. Darwin's theory thus makes the testable prediction that whenever we look beyond the human scale, our evolved intuition should break down.

We have repeatedly tested this prediction, and the results overwhelmingly support it: our intuition breaks down at high speeds, where time slows down; on small scales, where particles can be in two places at once; and at high temperatures, where colliding particles change identity. To me, an electron colliding with a positron and turning into a Z-boson feels about as intuitive as two colliding cars turning into a cruise ship. The point is that if we dismiss seemingly weird theories out of hand, we risk dismissing the correct theory of everything, whatever it may be.

If the mathematical universe hypothesis is true, then it is great news for science, allowing the possibility that an elegant unification of physics and mathematics will one day allow us to understand reality more deeply than most dreamed possible. Indeed, I think the mathematical cosmos is the best theory of everything that we could hope for: it would mean no aspect of reality is off-limits from our scientific quest to uncover regularities and make quantitative predictions.

However, it would also shift the ultimate question once again. We might abandon as misguided the question of which particular mathematical equations describe all of reality, and instead ask how to compute the frog's view of the universe – our observations – from the bird's view. That would determine whether we have uncovered the true structure of our universe, and help us figure out which corner of the mathematical cosmos is our home. ■

RANDOM REALITY

Does chance rule the cosmos – and if so, what does it mean for us, asks **Michael Brooks**

"OH, I am fortune's fool," says Romeo. Rest easy, lover boy; we all are. Or are we?

Romeo, having killed Tybalt and realising he must leave Verona or risk death, was expressing a view common in Shakespeare's time: that we are all marionettes, with some higher cause pulling the strings. Chance – let alone our own decision-making – plays little part in the unravelling of cosmic designs.

Even processes that inherently involved chance were pre-determined. Long before dice were used for gaming, they were used for divination. Ancient thinkers thought the gods determined the outcome of a die roll; the apparent randomness resulted from our ignorance of divine intentions.

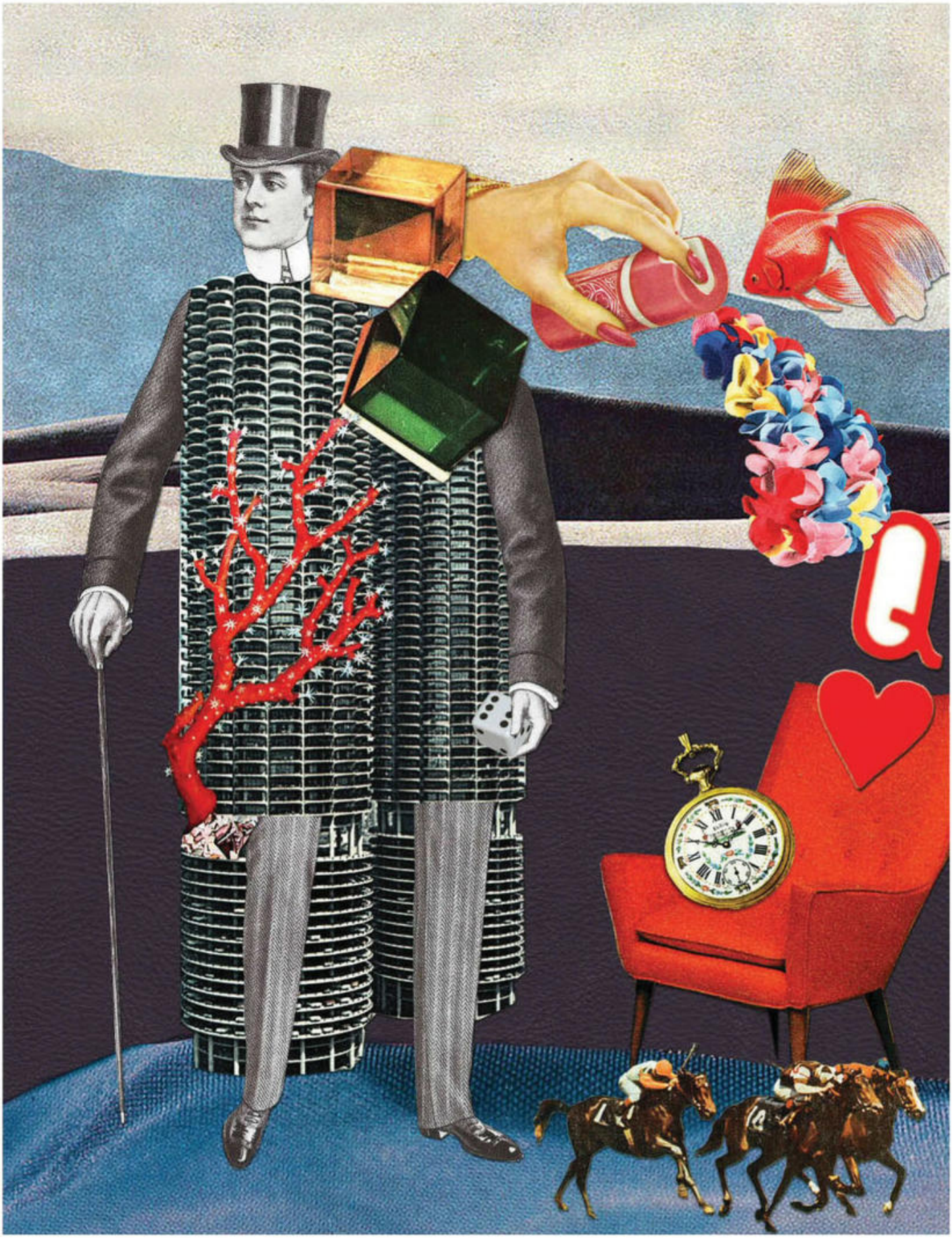
Oddly, modern science at first did little to change that view. Isaac Newton devised laws of motion and gravitation that connected everything in the cosmos with a mechanism run by a heavenly hand. The motion of the stars and planets followed the same strict laws as a cart pulled by a donkey. In this clockwork universe, every effect had a traceable cause.

If Newton's universe left little room for randomness, it did at least provide tools to second-guess the Almighty's intentions. If you had all the relevant facts pertaining to a die roll at your fingertips – trajectory, speed, roughness of the surface and so on – you could, in theory, calculate which face would end up on top. In practice this is far too complex a task. But it showed that randomness was nothing intrinsic; just a reflection of our lack of information.

Confidence in cosmic predictability led the French mathematician and physicist Pierre-Simon de Laplace to assert, a century after Newton, that a sufficiently informed intelligence could forecast everything that is going to happen in the universe – and, working backwards, tell you everything that did happen, right back to the cosmic beginnings.

It's a glorious and rather discomfiting idea. If everything really is predictable, then surely all is pre-determined and free will is an illusion? Romeo, in other words, is right. Perhaps so, says physicist Valerio Scarani, who studies randomness and its limits at the Centre for Quantum Technologies in Singapore. "One may believe that a single causal chain determines everything – call it God, the big bang or robots-behind-the-matrix," he says. "Then there is no randomness."

The connection between a universe that admits randomness and one that admits free will is an old one, says Scarani. The 13th century Christian ›



EUGENIA LOLI

STEVE MARCUS/REUTERS

philosopher Thomas Aquinas insisted a perfect universe must contain randomness to allow humans their autonomy. But it was also there to limit them. God made humans with less than divine abilities, so there must be a sphere of events beyond our control.

It wasn't until about two centuries after Newton that anyone began seriously to challenge the notion of a predictable cosmos. In 1859, Scottish physicist James Clerk Maxwell drew attention to the huge disparities in outcome that can stem from tiny factors affecting the collisions of molecules.

This was the beginnings of chaos theory. In its most familiar guise of the butterfly effect – that the flap of a butterfly's wings in Brazil might set off a tornado in Texas, as the chaos theorist Edward Lorenz put it in 1972 – this seems to restore unpredictability to the world. With a sufficiently complex system, even the tiniest approximation while working at the limits of your clock, barometer or ruler, or the slightest rounding error in a computation, can drastically affect the result. This is what makes the weather so hard to predict (see "The weather man", right). Its eventual state is highly dependent on the initial measurement – and we can never have a perfect initial measurement.

So, small, human-scale decisions might indeed matter on the wider stage. Romeo's predicament traces back to the initial conditions that first put him in the same room as Juliet – or further back still. Take that too far, though, and we might trace them back to before our ancestors came down from the trees, which seems to circumvent any sensible notion of human free will.

It's a head-scratcher, alright – but as yet we are only scratching the surface.

Because while we seem to occupy a reality where causes lead to predictable effects, dig down and that's apparently not how things work at all. Quantum theory, developed in stages since the early 20th century, is our working theory of reality at its most basic – and it does away with cast-iron certainty entirely. "It appears to us, via quantum experiments, that nature is fundamentally random," says Adrian Kent, a mathematician at the University of Cambridge.

"DIG DEEP DOWN INTO REALITY, AND IT SEEMS CAUSES DON'T LEAD TO PREDICTABLE EFFECTS"

Fire a single photon of light at a half-silvered mirror, and it might pass through or be reflected: quantum rules give us no way to tell beforehand. Give an electron a choice of two slits in a wall to pass through, and it chooses at random. Wait for a single radioactive atom to emit a particle, and you might wait a millisecond or a century. This rather lackadaisical attitude to classical certainties could even account for why we are here in the first place. A quantum vacuum containing nothing can randomly and spontaneously generate something. Such a careless energy fluctuation might best explain how our universe began.

Explaining the explanation is trickier. We don't know where the quantum rules came from; all we know is that the mathematics behind them, rooted in uncertainty, corresponds to reality observed up close. That starts with the Schrödinger equation, which describes how a quantum particle's properties evolve over time. An electron's position, for example, is given by an "amplitude" smeared over space, and there is a set of mathematical rules you can apply to find the probability that any particular measurement will pinpoint the electron to any particular position.

That's no guarantee the electron will be in that position at any one time. But by repeatedly doing the same measurement, resetting the system each time, the distribution of results will match the Schrödinger equation's predictions. The repeated, predictable patterns of the classical world are ultimately the result of many unpredictable processes.

The repercussions are interesting. Say you want to walk through a wall; quantum theory says it's possible. Each one of your atoms has a position that could – randomly – turn out to be on the other side of the wall when it interacts. That event's probability is exceedingly low, and the probability that all of your atoms will simultaneously locate to the other side of the wall is infinitesimally

"IF RANDOMNESS PROVES TO BE AN ILLUSION, FREE WILL MIGHT BE TOO"

small. A nasty bruise is the sum of all the other probabilities. Welcome to reality.

Einstein was particularly exercised by this probabilistic approach to real-world events, famously complaining it was akin to God playing dice. He conjectured that there must be some missing information that would tell you the measurement's outcome in advance.

Hidden realities

In 1964, the physicist John Bell laid out a way to test for such "hidden variables". His idea has since been implemented time and again, mainly using entangled pairs of photons. Entangled particles are a staple feature of the quantum world. They have interacted at some point in the past and now appear to have shared properties, such that a measurement on particle A will instantaneously affect what you get from a measurement on particle B, and vice versa.

What's behind this? The details of Bell's tests are complex and subtle, but the principle is akin to a sport in which two groups of experimenters play according to different rules. Team Alpha assumes that the quantum correlations are down to some hidden exchange of information, and make measurements accordingly. Team Beta, in contrast, assumes the correlations materialise at random on measurement.

And Team Beta wins every time. The weird correlations of the quantum world derive from fundamental randomness.

Or do they? Physicists are still investigating loopholes in the way we do quantum measurements that might skew the results and simulate randomness – the fact that we can't measure the state of photons with 100 per cent accuracy, for instance, or even the question of whether we have free will in choosing the measurements we make. "I think it's premature to say we've closed all the important Bell loopholes," says Kent.

It is possible that quantum theory's vagaries might one day be explained, perhaps by compromising some other cherished principle, such as Einstein's relativity. Or maybe someone will come up with some more intuitive, non-random theory that reproduces all the predictions of quantum theory and makes some stronger ones as well. "That hypothetical theory would be a new theory – a successor to quantum theory, not a version of it," Kent says.

Terry Rudolph, a physicist at Imperial College London, agrees. Quantum theory is our ultimate theory of nature, and it seems to suggest the universe is random, but that is no guarantee it is. "I don't think we can ever prove it," he says.

If so, randomness might still prove to be an illusion – and with it, perhaps our free will. "Then quantum physics is just part of the big conspiracy," says Scarani.

Fortune's fools? Perhaps we're not at liberty to decide. ■

MICHAEL ZUMSTEIN/AGENCE VU/CAMERA PRESS

THE WEATHER MAN
KEN MYLNE

Head of weather science numerical modelling at the UK Met Office

How do you forecast the weather?
We set up a model to represent the current state of the atmosphere based on many observations. From that, the model projects forward in time and calculates how the atmosphere may evolve. The outcome of the forecast is very sensitive to small errors in the initial state, so we run what we call an ensemble forecast. Instead of just running the model once, we make a series of small changes to the initial state and re-run the model a large number of times to get a set of forecasts. On some days the model runs may be similar, which gives us a high level of confidence in the forecast; on other days, the model runs can differ radically so we have to be more cautious.

How certain can you be about forecasts?
The level of confidence varies from day to day and from forecast to forecast. In some circumstances you can get big differences between the forecasts in the ensemble. The biggest uncertainties are often around big storms and the dramatic weather everyone cares about, because the atmosphere has to be in a sensitive, unstable state to generate that high-impact weather. The chaotic nature of the atmospheric system does impose fundamental limits on predictability. In terms of day-to-day weather, that limit is typically between 10 days and two weeks using probabilistic forecasts.

From 2011, the Met Office started presenting rain forecasts using probabilities. Was that controversial?
We'd been debating it for a long time. The Americans have been putting out probability of precipitation forecasts for many years, and it's quite accepted there. The argument in favour is that often you cannot – for good scientific reasons – say definitely that it will or will not be raining. So you are giving people much better information if you tell them the probability of rainfall. While we recognise that some people find probabilities difficult to understand, lots of people do understand them and make better decisions as a result.

MATT MURPHY

In two minds

Probability rules the quantum world – or is it just you that's uncertain, asks **Matthew Chalmers**

SNATCH a toy from the tiniest of infants, and the reaction is likely to disappoint you. Most seem to conclude that the object has simply ceased to exist. This rapidly changes. Within the first year or so, playing peekaboo also becomes fun. As babies, we soon grasp that stuff persists unchanged even when we are not looking at it.

Granted, at that age we know nothing of quantum theory. In the standard telling, this most well-tested of physical theories – fount of the computers, lasers and cellphones that our adult souls delight in – informs us that reality's basic building blocks take on a very different, nebulous form when no one is looking. Electrons, quarks or entire atoms can easily be in two different places at once, or have many properties simultaneously. We cannot predict with certainty which of the many possibilities we will see: that is all down to the random hand of probability.

That's not the way our grown-up, classical world seems to work, and physicists have been scrabbling around for the best part of a century to explain the puzzling mismatch. To no avail. Faced with reality at its most fundamental, we end up babbling baby talk again.

David Mermin thinks he has something sensible to say. An atomic physicist at Cornell University in Ithaca, New York, he has spent most of his half-century-long career rejecting philosophical musings about the nature of quantum theory. Now he's had an epiphany. The way to solve our quantum conundrums is to abandon the ingrained idea that we can ever achieve an objective view of reality. According to this provocative idea, the world is not uncertain – we are.

The idea that an objective, universally valid view of the world can be achieved by making properly controlled measurements is perhaps the most basic assumption of modern science. It works well enough in the macroscopic, classical world. Kick a football, and Newton's laws of motion tell you where it will be later, regardless of who is watching it and how.

Kick a quantum particle such as an electron or a quark, though, and the certainty vanishes. At best, quantum theory allows you to calculate the probability of one outcome from many encoded in a multifaceted wave function that describes the particle's state. Another observer making an identical measurement on an identical particle might measure something very different. You have no way of saying for sure what will happen.

So what state is a quantum object in when no one is looking? The most widely accepted answer is the Copenhagen interpretation, so named after the site of many early quantum musings. Schrödinger's notorious cat illustrates its conclusion. Shut in a box with a vial of lethal gas that might, or might not, have been released by a random quantum event such as a radioactive decay, the unfortunate feline hangs in limbo, both alive and dead. Only when you open the box does the cat's wave function "collapse" from its multiple possible states into a single real one.

This opens a physical and philosophical can of worms. Einstein pointedly asked whether the observations of a mouse would be sufficient to collapse a wave function. If not, what is so special about human consciousness? If our measurements truly do affect reality, that also opens the door to effects such as "spooky action at a distance" – Einstein's dismissive phrase to describe how observing a wave function can seemingly collapse another one simultaneously on the other side of the universe. ›

LOST IN SPACE

From a human perspective, physics has a problem with time. We have no difficulty defining a special moment called "now" that is distinct from the past and the future, but our theories cannot capture the essence of the moment. The laws of nature deal only with what happens between certain time intervals.

David Mermin of Cornell University claims to have solved this problem using a principle similar to the one he and others have applied to quantum theory (see main story). We should simply abandon the notion that an objectively determinable space-time exists.

Instead of forming a series of slices or layers that from some viewpoint correspond to a "now" or "then", Mermin's space-time is a mesh of intersecting filaments relating to the experiences of different people. "Why promote space-time from a 4D diagram, which is a useful conceptual device, to a real essence?" he asks. "By identifying my abstract system with an objective reality, I fool myself into regarding it as the arena in which I live my life."

Things such as an interval of time or the dimensionality of space, after all, are not stamped on nature for us to read off; a newborn baby has no conception of them. They are merely useful abstractions we develop to account for what clocks and rulers do. Some of these high-level abstractions we construct for ourselves as we grow up, others were constructed by geniuses and have been passed on to us in school or in books, says Mermin. "And some of them, like quantum states, most of us never learn at all."

Then there is the mystery of how atoms and particles can apparently adopt split personalities, but macroscopic objects such as cats clearly can't, despite being made up of atoms and particles. Schrödinger's intention in introducing his cat was to highlight this inexplicable division between the quantum and classical worlds. The split is not only there, but also "shifty", in the words of quantum theorist John Bell: physicists contrive to put ever-larger objects into fuzzy quantum states, for instance – so we have no set way of defining where the boundary lies.

The Copenhagen interpretation simply ignores these quantum mysteries, famously leading Mermin to dub it the "shut up and calculate" approach in an article he wrote in 1989. He counted himself as an adherent. Although alternatives did exist – such as the many worlds interpretation, which suggests the universe divides into different paths every time anything is observed – none quite seemed to crack the central mystery.

Frequently wrong

Now Mermin thinks one does. It is not his idea: in fact, he spent more than a decade arguing against it with its originators, Carlton Caves of the University of New Mexico in Albuquerque, Christopher Fuchs of the Perimeter Institute for Theoretical Physics in Waterloo, Canada, and Rüdiger Schack of Royal Holloway, University of London.

Known as quantum Bayesianism, its ideas stem from reassessing the meaning of the wave function probabilities that seemingly govern the quantum world (see diagram, right). Conventionally, these are viewed as "frequentist" probabilities. In the same way that you might count up many instances of a coin falling heads or tails to conclude that the odds are 50/50, many measurements of a quantum system tell you the relative frequency of its multiple states cropping up.

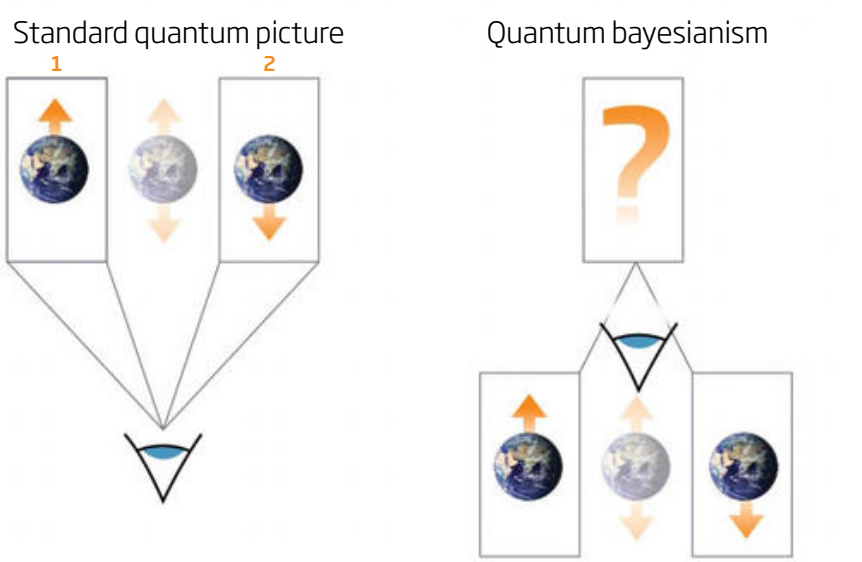
Despite its limitations, not least when dealing with single, isolated events, frequentist probability is popular throughout science for the way it turns an observer into an entirely objective counting machine. But an alternative, older approach to probability was devised by English clergyman Thomas Bayes in the 18th century. This is the sort of probability that crops up in a statement such as "there's a 40 per cent chance of rain today". Its value is not objective or fixed, but a fluid assessment based on many changing factors, such as current air pressure and how similar weather systems developed in the past. Acquire a new piece of information – see a bank of threatening cloud when you open the curtains in the morning, for example – and you might well update your prognosis to a 90 or 100 per cent chance of rain. The actual likelihood of rain has not changed; but your state of knowledge about it has.

Under what circumstances it is legitimate to use this more subjective type of probability is the subject of heated debate among mathematicians (see "Probability wars", page 56). The central argument of quantum Bayesianism, or QBism, is that, by applying it to the quantum world, whole new vistas open up. Measure the spin of an invisible electron, say, and you acquire new knowledge, and update your assessment of the probabilities accordingly, from uncertain to certain. Nothing needs to have changed at the quantum level. Quantum states, wave functions and all the other probabilistic apparatus of quantum mechanics do not represent objective truths about stuff in the real world. Instead, they are subjective tools that we use to organise our uncertainty about a measurement before we perform it. In other words: quantum weirdness is all in the mind. "It really is that simple," says Mermin.

Mind you, it took six weeks of intense discussions with Fuchs and Schack in South Africa in 2013 to finally convince Mermin that he had been a QBist all along. In November of that year, they published their conclusions together.

Uncertain uncertainty

According to quantum Bayesianism, quantum fuzziness just reflects our lack of knowledge of the world



Objects in the quantum world exist in a fuzzy combination of states. The act of measuring forces them to adopt a specific state (1 or 2)

The quantum states are all in our minds – they are just a fluid tool we use to understand our variable experiences of the world

For Mermin, the beauty of the idea is that the paradoxes that plague quantum mechanics simply vanish. Measurements do not "cause" things to happen in the real world, whatever that is; they cause things to happen in our heads. Spooky action at a distance is an illusion too. The appearance of a spontaneous change is just the result of two parties independently performing measurements that update their state of knowledge.

As for that shifty split, the "classical" world is where acts of measurements are continuous, because we see things with our own eyes. The microscopic "quantum" world, meanwhile, is where we need an explicit act of measurement with an appropriate piece of equipment to gain information. To predict outcomes in this instance, we require a theory that can take account of all the things that might be going on when we are not looking. For a QBist, the quantum-classical boundary is the split between what is going on in the real world and your subjective experience of it.

End of observers

Quantum theorist William Wootters of Williams College in Williamstown, Massachusetts, thinks this is the most exciting interpretation of quantum theory to have emerged in years, and points to historical precedents. "It addresses Schrödinger's concern that our own subjective experience has been explicitly excluded from physical science, and both requires and provides a place for the experiencing subject," he says.

Others are less keen. Carlo Rovelli of Aix-Marseille University in France proposed a similar, less extreme, observer-dependent idea called relational quantum mechanics in 1996. He worries that QBism relies too much on a philosophy espoused by German philosopher Immanuel Kant in the 18th century – that there is no direct experience of things, only that which we construct in our minds from sensory inputs. "I would prefer an interpretation of quantum theory that would make sense even if there were no humans to observe anything," he says.

Antony Valentini of Clemson University in South Carolina also thinks it moves things in the wrong direction. He paints a picture of someone setting up equipment to measure the energy of a particle, and then going off for a cup of tea. During the tea break, did the pointer on the equipment's dial have no definite orientation? A QBist would say maybe not, you can't tell – even though experience tells us a macroscopic object such as a pointer does always have a definite orientation. That view can't be taken seriously, says Valentini. "A physical theory should try to describe the physical world, not just some body of talk."

Schack counters that there is only one world out there, and we must find a way of unifying our classical and quantum interpretations of it – even if it means accepting we have no objective connection to reality in either sphere. "QBism abandons the idea that nature can be described adequately from the perspective of a detached observer," he says.

For him the strongest sign that QBism is on the right track is a thought experiment called Wigner's friend. Imagine you are standing outside a closed room where a friend is about to open the box containing Schrödinger's cat. Your friend witnesses a clear outcome: the cat

"The beauty of the idea is that the paradoxes of quantum theory just vanish"

is either alive or dead. But you must assign a set of probabilities based on a superposition of all the possible states of the cat and the reports your friend might make of it. Who's right? Both, say QBists: there is no paradox if a measurement outcome is always personal to the person experiencing it.

With all the zeal of a convert, Mermin has recently sought to convince detractors by applying QBist reasoning to the problems of an entity that has nothing to do with quantum theory, and nothing to do with probability: space-time (see "Lost in space", left).

But Caslav Brukner of the University of Vienna in Austria wonders how far such approaches can take us. "I do not see in QBism the power to explain why quantum theory has the very mathematical and conceptual structure it does," he says. Other theories about the world at its most fundamental could have similar Bayesian underpinnings – so why specifically does quantum theory come up with the right answers? Like many who have inspected the undercarriage of quantum mechanics, Brukner would prefer to reconstruct it from a core set of principles or axioms.

You might wonder whether all this matters, given that quantum theory does such a stupendous job of describing the world and supplying us with technological innovation. That is true up to a point, says Rovelli – but our lack of intuitive understanding hampers our search for some greater theory that can embrace all of physics from the smallest to the largest scales. "If we want to better understand the world, for instance, for quantum gravity or for cosmology, it does matter," he says.

Faced with the prospect of abandoning scientific objectivity, the temptation to shut up and calculate might be as strong as ever. But perhaps quantum Bayesianism provides a way to have our cake and eat it. Shifting quantum theory's weirdness into our own minds doesn't diminish our power to calculate with it – but might just make us shut up about how shocking it all is. ■

Spaghetti functions

What possessed an architect to boil down the beauty of pasta to a few bare mathematical formulae, asks Richard Webb

ALPHABETTI spaghetti: now there was a name to conjure with when I was a kid. Succulent little pieces of pasta, each shaped into a letter of the alphabet, served up in a can with lashings of tomato sauce. Delicious, nutritious – and best of all they made playing with your food undeniably educational.

A few decades on, in an upscale Italian restaurant near the London offices of *New Scientist*, I decide against sharing this reminiscence of family mealtimes with my lunching partner. George Legendre doesn't look quite the type. For one thing, he is French, and possibly indisposed to look kindly on British culinary foibles. For another, he is an architect, designer and connoisseur of all things pasta who in 2011 compiled the first comprehensive mathematical taxonomy of the stuff.

According to some measures, pasta is now the world's favourite food. Something like 13 million tonnes are produced annually around the globe, with Italy topping the league of both producers and consumers, according to figures from the International Pasta Organisation, a trade body. The average Italian gets through 26 kilograms – that's the uncooked mass – of pasta each year.

The plate of *paccheri* in front of me seems positively modest by comparison. To my untrained eye, it consists of large, floppy and slightly misshapen penne. I might not be too wide of the mark. "If you look carefully, there are probably only three basic topological shapes in pasta – cylinders, spheres and ribbons," Legendre says.

Nevertheless, that simplicity has, in the hands of pasta maestros throughout the world, spawned a multiplicity of complex forms – and inspired many a designer before Legendre (see "Primi piatti", page 116). It was a late-night glass of wine too many at his architectural practice in London that inspired Legendre, together with his colleague Jean-Aimé Shu, into using mathematics to bring order to this chaotic world.

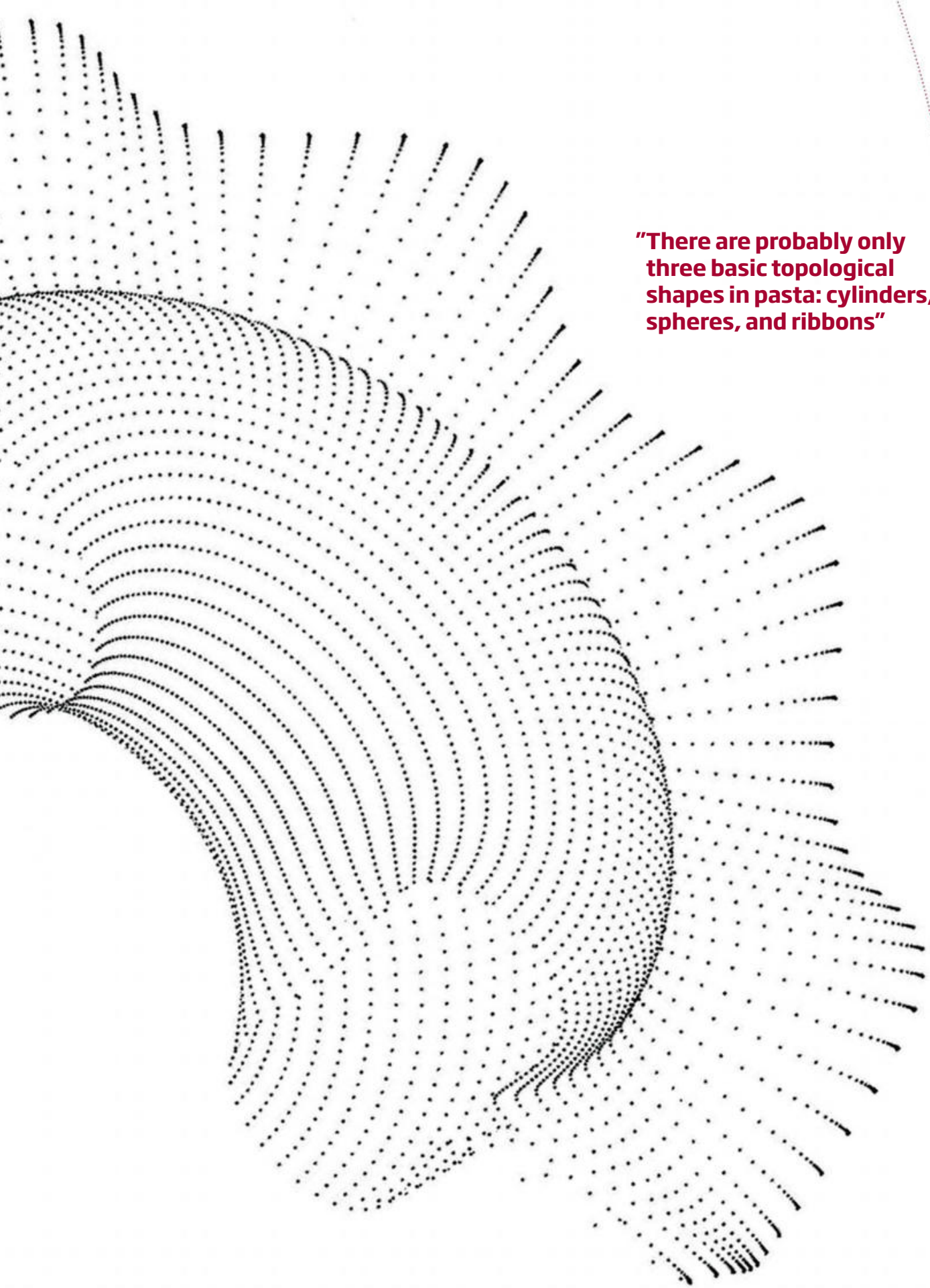
"The first thing we did was order lots of pasta," Legendre says. Then, using their design know-how, they set about modelling every shape they could lay their hands on to derive formulae that encapsulate their forms. "It took almost a year and almost bankrupted the company," he says.

For each shape, they needed three expressions, each describing its form in one of the three dimensions. This provides a set of coordinates that, plotted on a graph, faithfully represents the pasta's 3D shape. The curvaceous shapes of most pasta lend themselves to mathematical representions mainly through oscillating sine and cosine functions.

For some pastas, the right recipe was obvious. Spaghetti, for example, is little more than an extruded circle. The sine and cosine of a single angle serves to define the coordinates of the points enclosing its unvarying cross section, and a simple constant characterises its length. Similarly, grain-like *puntalette* are just deformed spheres. The sines and cosines of two angles, together with different multiplying factors to stretch the shape out in three dimensions, provide the necessary expressions. "The compactness of the expression is beautiful," says Legendre.

Other shapes were harder to crack. ›

STEFANO GRAZIANI



"There are probably only three basic topological shapes in pasta: cylinders, spheres, and ribbons"

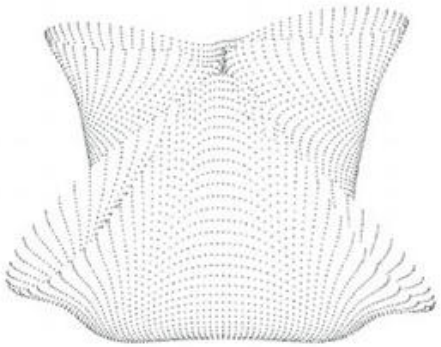
PRIMI PIATTI

The Italian designer Giorgetto Giugiaro has a string of supercars to his name, conceived for the likes of Ferrari, Maserati and Lamborghini. In 1999 he was voted "car designer of the century" by an international jury of motoring journalists.

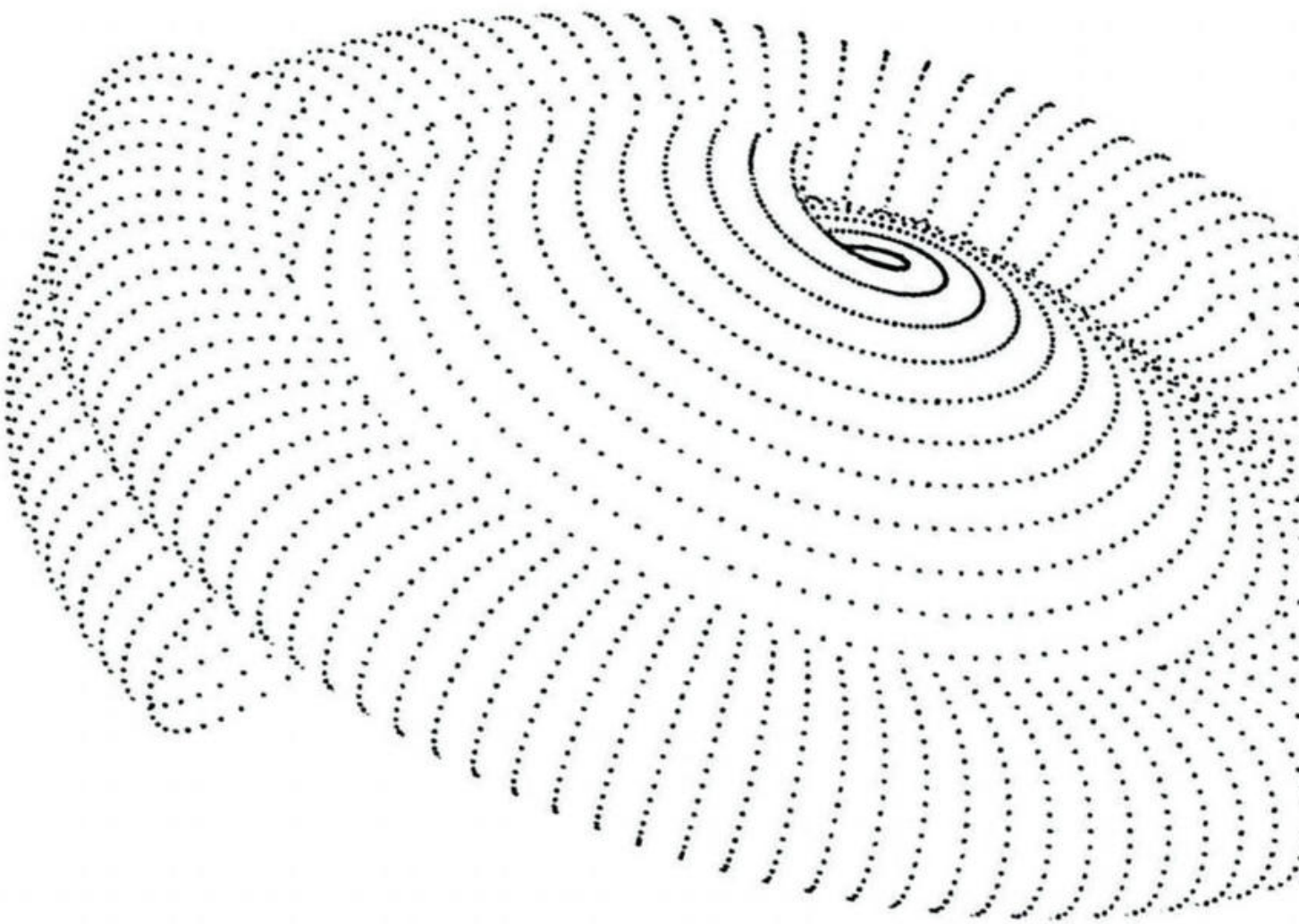
Less well known are his activities as a designer of pasta. In 1983, the Neapolitan manufacturer Voiello commissioned him to design a new shape compatible with the traditional manufacturing method of extrusion, in which the pasta dough is forced through a slit in a bronze die. In the event, his "Marille" design, consisting of two parallel tubes with a flap protruding from their join, rather landed him in hot water. While pleasing on the eye, its intricacy meant that different parts of the pasta cooked at vastly different rates.

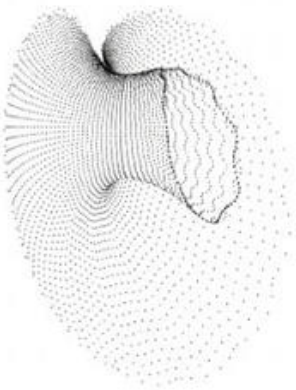
In 1987, the celebrated designer Philippe Starck conceived a similar-looking shape for the French pasta maker Panzani. Called the Mandala, it resembled a yin-yang symbol elongated in a third dimension. It, too, failed to break through into the pasta big time.

Fun rather than practicality seemed to be on the minds of two designers from the Bezalel Academy of Arts and Design in Jerusalem, Israel, who devised their own pasta in 2009. Resembling penne, it could be used as a whistle before cooking.



"The pasta taxonomy proves that immense complexity has simple beginnings"





Scrunched-up *saccottini*, for example, look for all the world like the crocheted representation of a hyperbolic plane that adorns my desk, and its shape is captured by a complex mathematical mould of multiplied sines and cosines. Simple features such as the slanted ends of *penne* take some low modelling cunning, involving chopping the pasta into pieces, each represented by slightly different equations.

Sharp inflections, such as the undulating crests of the cockscomb-like *galletti*, proved to be tricky too, though trigonometric functions again turn out to be the best tools for the job: raising sines and cosines to a higher power constricts the smooth, oscillating shape of the function into something approaching a spike. A similar technique can be used to broaden out the function into something approaching a right angle – a trick Legendre dubs an "asymptotic box". "Saying to colleagues you're developing mathematics to make a box makes them think you're crazy," he says.

In the end, he had a compendium of 92 pasta shapes, each exactly modelled and divided into categories according to the mathematical relationships revealed between them – some obvious, some less so. The twisted ribbons of *sagne incannulate* and the "little hats", *cappelletti*, turn out to be topologically identical: given sufficiently pliant dough, deft hands could stretch, twist and remould one shape into the other without the intervention of a knife or pair of scissors.

Whimsical though such insights may be, the project has a serious note too. Legendre's pasta taxonomy provides a playful proof that immense variety and seeming complexity can be reduced to simple mathematical beginnings. Legendre is convinced that could lead to a new, more efficient way of translating design into engineering, useful for structures on a much larger scale. Plans for an arbitrarily complex skyscraper, for example, might be reduced to equations for each of its three dimensions just like those that define the pasta shapes. "You can see the equations for cross section as indicative of a floor, with a third equation for the elevation," he says.

In fact, he has already put the principle into practice. Legendre's Henderson Waves bridge in Singapore has an undulating form more than a little reminiscent of graceful pasta-like curves, and was modelled using exactly the same principles. "I just gave the engineers equations," he says.

His own pasta shape is next on the menu. His original intention there was to bridge a gap between his passion and his profession: the relative dearth in the pasta world of the sturdy, rectilinear shapes that form the basis of most architecture. In the current pasta taxonomy, this sort of form is represented only by *trenne*, hollow bars with a triangular cross section. But making such seemingly basic shapes accurately turns out to be fiendishly difficult using the traditional process of extruding the dough through a bronze die – a wrinkle that Legendre is currently trying to iron out with pasta manufacturer.

Do things need to be that complex? My imagination is piqued by the idea that I too might one day hook my computer, equipped with a pasta modelling package, to a 3D printer and print my own pasta. But Legendre is not so sure the results would tickle my taste buds. Each pasta shape is the product of a different regional or local tradition, and centuries of painstaking R&D to match the right shape with the right sauce, he says.

That's the kind of love mathematics cannot buy – but it might, perhaps, be food for another project. "I would love to see a book that deals with the right seasoning as rigorously," he says wistfully.

Me too: perhaps then alphabetti spaghetti and its oozing tomato sauce will be given the belated recognition it deserves. Meanwhile, I have to admit I'm quite enjoying the melange of buffalo mozzarella, aubergines and tomatoes in front of me. Legendre, for his part, is having the risotto. ■

The right set of coordinates faithfully reproduces any pasta

Alice's secrets in Wonderland

There's a hidden mathematical meaning behind Lewis Carroll's classic tale, says **Melanie Bayley**

WHAT would Lewis Carroll's *Alice's Adventures in Wonderland* be without the Cheshire Cat, the trial, the Duchess's baby or the Mad Hatter's tea party? Look at the original story that the author told Alice Liddell and her two sisters one day during a boat trip near Oxford, though, and you will find that these famous characters and scenes are missing from the text.

As I embarked a few years back on my DPhil investigating Victorian literature, I wanted to know what inspired these later additions. The critical literature focused mainly on Freudian interpretations of the book as a wild descent into the dark world of the subconscious. There was no detailed analysis of the added scenes, but from the mass of literary papers, one stood out: in 1984 Helena Pycior of the University of Wisconsin-Milwaukee had linked the trial of the Knave of Hearts with a Victorian book on algebra. Given the author's day job, it was somewhat surprising to find few other reviews of his work from a mathematical perspective. Carroll was a pseudonym: his real name was Charles Dodgson, and he was a mathematician at Christ Church College, Oxford.

The 19th century was a turbulent time for mathematics, with many new and controversial concepts, like imaginary numbers, becoming widely accepted in the mathematical community. Putting *Alice's Adventures in Wonderland* in this context, it becomes clear that Dodgson, a stubbornly conservative mathematician, used some of the missing scenes to satirise these radical new ideas.

Even Dodgson's keenest admirers would admit he was a cautious mathematician who produced little original work. He was, however, a conscientious tutor, and, above everything, he valued the ancient Greek textbook Euclid's *Elements* as the epitome of mathematical thinking. Broadly speaking, it covered the geometry of circles, quadrilaterals, parallel lines and some basic trigonometry. But what's really striking about *Elements* is its rigorous reasoning: it starts with a few incontrovertible truths, or axioms, and builds up complex arguments through simple, logical steps. Each proposition is stated, proved and finally signed off with QED.

For centuries, this approach had been seen as the pinnacle of mathematical and logical reasoning. Yet to Dodgson's dismay, contemporary mathematicians weren't always as rigorous as Euclid. He dismissed

IMAGINARY MATHEMATICS

The real numbers, which include fractions and irrational numbers like π that can nevertheless be represented as a point on a number line, are only one of many number systems.

Complex numbers, for example, consist of two terms – a real component and an "imaginary" component formed of some multiple of the square root of -1, now represented by the symbol *i*. They are written in the form a + b*i*.

The Victorian mathematician William Rowan Hamilton took this one step further, adding two more terms to make quaternions, which take the form a + b*i* + c*j* + d*k* and have their own strange rules of arithmetic.

ANDREW HEM

their writing as "semi-colloquial" and even "semi-logical". Worse still for Dodgson, this new mathematics departed from the physical reality that had grounded Euclid's works.

By now, scholars had started routinely using seemingly nonsensical concepts such as imaginary numbers – the square root of a negative number – which don't represent physical quantities in the same way that whole numbers or fractions do. No Victorian embraced these new concepts wholeheartedly, and all struggled to find a philosophical framework that would accommodate them. But they gave mathematicians a freedom to explore new ideas, and some were prepared to go along with these strange concepts as long as they were manipulated using a consistent framework of operations. To Dodgson, though, the new mathematics was absurd, and while he accepted it might be interesting to an advanced mathematician, he believed it would be impossible to teach to an undergraduate.

Outgunned in the specialist press, Dodgson took his mathematics to his fiction. Using a technique familiar from Euclid's proofs, reductio ad absurdum, he picked apart the "semi-logic" of the new abstract mathematics, mocking its weakness by taking these premises to their logical conclusions, with mad results. The outcome is *Alice's Adventures in Wonderland*.

Algebra and hookahs

Take the chapter "Advice from a caterpillar", for example. By this point, Alice has fallen down a rabbit hole and eaten a cake that has shrunk her to a height of just 3 inches. Enter the Caterpillar, smoking a hookah pipe, who shows Alice a mushroom that can restore her to her proper size. The snag, of course, is that one side of the mushroom stretches her neck, while another shrinks her torso. She must eat exactly the right balance to regain her proper size and proportions.

While some have argued that this scene, with its hookah and "magic mushroom", is about drugs, I believe it's actually about what Dodgson saw as the absurdity of symbolic algebra, which severed the link between algebra, arithmetic and his beloved geometry. Whereas the book's later chapters contain more specific mathematical analogies, this scene is subtle and playful, setting the tone for the madness that will follow.

The first clue may be in the pipe itself: the word "hookah" is, after all, of Arabic origin, like "algebra", and it is perhaps striking that Augustus De Morgan, the first British ›

mathematician to lay out a consistent set of rules for symbolic algebra, uses the original Arabic translation in *Trigonometry and Double Algebra,* which was published in 1849. He calls it "al jebr e al mokabala" or "restoration and reduction" – which almost exactly describes Alice's experience. Restoration was what brought Alice to the mushroom: she was looking for something to eat or drink to "grow to my right size again", and reduction was what actually happened when she ate some: she shrank so rapidly that her chin hit her foot.

De Morgan's work explained the departure from universal arithmetic – where algebraic symbols stand for specific numbers rooted in a physical quantity – to that of symbolic algebra, where any "absurd" operations involving negative and impossible solutions are allowed, provided they follow an internal logic. Symbolic algebra is essentially what we use today as a finely honed language for communicating the relations between mathematical objects, but Victorians viewed algebra very differently. Even the early attempts at symbolic algebra retained an indirect relation to physical quantities.

De Morgan wanted to lose even this loose association with measurement, and proposed instead that symbolic algebra should be considered as a system of grammar. "Reduce" algebra from a universal arithmetic to a series of logical but purely symbolic operations, he said, and you will eventually be able to "restore" a more profound meaning to the system – though at this point he was unable to say exactly how.

When Alice loses her temper

The madness of Wonderland, I believe, reflects Dodgson's views on the dangers of this new symbolic algebra. Alice has moved from a rational world to a land where even numbers behave erratically. In the hallway, she tried to remember her multiplication tables, but they had slipped out of the base-10 number system we are used to. In the caterpillar scene, Dodgson's qualms are reflected in the way Alice's height fluctuates between 9 feet and 3 inches. Alice, bound by conventional arithmetic where a quantity such as size should be constant, finds this troubling: "Being so many different sizes in a day is very confusing," she complains. "It isn't," replies the Caterpillar, who lives in this absurd world.

The Caterpillar's warning, at the end of this scene, is perhaps one of the most telling clues to Dodgson's conservative mathematics. "Keep your temper," he announces. Alice presumes he's telling her not to get angry, but although he has been abrupt he has not been particularly irritable at this point, so it's a somewhat puzzling thing to announce. To intellectuals at the time, though, the word "temper" also retained its original sense of "the proportion in which qualities are mingled", a meaning that lives on today in phrases such as "justice tempered with mercy". So the Caterpillar could well be telling Alice to keep her body in proportion – no matter what her size.

This may again reflect Dodgson's love of Euclidean geometry, where absolute magnitude doesn't matter: what's important is the ratio of one length to another when considering the properties of a triangle, for example. To survive in Wonderland, Alice must act like a Euclidean geometer, keeping her ratios constant, even if her size changes.

Of course, she doesn't. She swallows a piece of mushroom and her neck grows like a serpent with predictably chaotic results – until she balances her shape with a piece from the other side of the mushroom. It's an important precursor to the next chapter, "Pig and pepper", where Dodgson parodies another type of geometry.

By this point, Alice has returned to her proper size and shape, but she shrinks herself down to enter a small house. There she finds the Duchess in her kitchen nursing her baby, while her Cook adds too much pepper to the soup, making everyone sneeze except the Cheshire Cat. But when the Duchess gives the baby to Alice, it somehow turns into a pig.

The target of this scene is projective geometry, which examines the properties of figures that stay the same even when the figure is projected onto another surface – imagine shining an image onto a moving screen and then tilting the screen through different angles to give a family of shapes. The field involved various notions that Dodgson would have found ridiculous, not least of which is the "principle of continuity".

Jean-Victor Poncelet, the French mathematician who set out the principle, describes it as follows: "Let a figure be conceived to undergo a certain continuous variation, and let some general property concerning it be granted as true, so long as the variation is confined within certain limits; then the same property will belong to all the successive states of the figure."

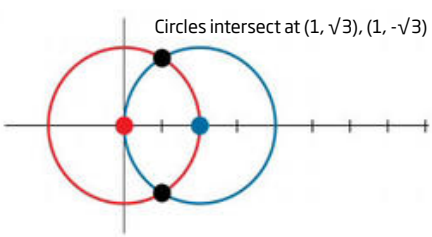
The case of two intersecting circles is perhaps the simplest example to consider. Solve their equations, and you will find that they intersect at two distinct points. According to the principle of continuity, any continuous transformation to these circles – moving their centres away from one another, for example – will preserve the basic property that they intersect at two points. It's just that when their centres are far enough apart the solution will involve an imaginary number that can't be understood physically (see diagram).

Of course, when Poncelet talks of "figures", he means geometric figures, but Dodgson playfully subjects Poncelet's "semi-colloquial" argument to strict logical analysis and takes it to its most extreme conclusion. What works for a triangle should also work for a baby; if not, something is wrong with the principle, QED. So Dodgson turns a baby into a pig through the principle of continuity. Importantly, the baby retains most of its original features, as any object going through a continuous transformation must. His limbs are still held out like a starfish, and he has a queer shape, turned-up nose and small eyes. Alice only realises he has changed when his

The continuity principle

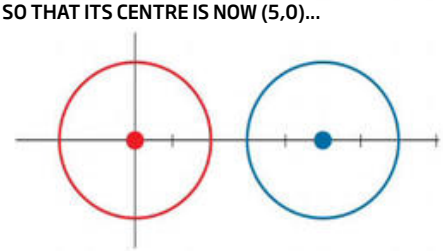
This principle, which so upset Charles Dodgson, stated that a mathematical figure should retain some of its original properties even under drastic transformations

CONSIDER TWO CIRCLES CENTRED ON (0,0) AND (2,0)



These circles intersect at two places, and under the principle of continuity you can assume that they will always intersect in two places, even if they move apart and are no longer touching!

SO IF THE BLUE CIRCLE MOVES SO THAT ITS CENTRE IS NOW (5,0)...



...these circles still intersect at two points $(\frac{5}{2}, \frac{3}{2}i)$ and $(\frac{5}{2}, -\frac{3}{2}i)$, where i is $\sqrt{-1}$

ANDREW HEM

sneezes turn to grunts.

The baby's discomfort with the whole process, and the Duchess's unconcealed violence, signpost Dodgson's virulent mistrust of "modern" projective geometry. Everyone in the pig and pepper scene is bad at doing their job. The Duchess is a bad aristocrat and an appallingly bad mother; the Cook is a bad cook who lets the kitchen fill with smoke, over-seasons the soup and eventually throws out her fire irons, pots and plates.

Alice, angry now at the strange turn of events, leaves the Duchess's house and wanders into the Mad Hatter's tea party, which explores the work of the Irish mathematician William Rowan Hamilton. Hamilton died in 1865, just after *Alice* was published, but by this time his discovery of quaternions in 1843 was being hailed as an important milestone in abstract algebra, since they allowed rotations to be calculated algebraically.

Just as complex numbers work with two terms, quaternions belong to a number system based on four terms (see "Imaginary mathematics", page 118). Hamilton spent years working with three terms – one for each dimension of space – but could only make them rotate in a plane. When he added the fourth, he got the three-dimensional rotation he was looking for, but he had trouble conceptualising what this extra term meant. Like most Victorians, he assumed this term had to mean something, so in the preface to his *Lectures on Quaternions* of 1853 he added a footnote: "It seemed (and still seems) to me natural to connect this extra-spatial unit with the conception of time."

Where geometry allowed the exploration of space, Hamilton believed, algebra allowed the investigation of "pure time", a rather esoteric concept he had derived from Immanuel Kant that was meant to be a kind of Platonic ideal of time, distinct from the real time we humans experience. Other mathematicians were polite but cautious about this notion, believing pure time was a step too far.

The parallels between Hamilton's maths and the Hatter's tea party – or perhaps it should read "t-party" – are uncanny. Alice is now at a table with three strange characters:

"Wonderland's madness reflects Carroll's views on the dangers of the new symbolic algebra"

the Hatter, the March Hare and the Dormouse. The character Time, who has fallen out with the Hatter, is absent, and out of pique he won't let the Hatter move the clocks past six.

Reading this scene with Hamilton's maths in mind, the members of the Hatter's tea party represent three terms of a quaternion, in which the all-important fourth term, time, is missing. Without Time, we are told, the characters are stuck at the tea table, constantly moving round to find clean cups and saucers.

Their movement around the table is reminiscent of Hamilton's early attempts to calculate motion, which was limited to rotatations in a plane before he added time to the mix. Even when Alice joins the party, she can't stop the Hatter, the Hare and the Dormouse shuffling round the table, because she's not an extra-spatial unit like Time.

The Hatter's nonsensical riddle in this scene – "Why is a raven like a writing desk?" – may more specifically target the theory of pure time. In the realm of pure time, Hamilton claimed, cause and effect are no longer linked, and the madness of the Hatter's unanswerable question may reflect this.

Alice's ensuing attempt to solve the riddle pokes fun at another aspect of quaternions: their multiplication is non-commutative, meaning that $x \times y$ is not the same as $y \times x$. Alice's answers are equally non-commutative. When the Hare tells her to "say what she means", she replies that she does, "at least I mean what I say – that's the same thing". "Not the same thing a bit!" says the Hatter. "Why, you might just as well say that 'I see what I eat' is the same thing as 'I eat what I see'!"

It's an idea that must have grated on a conservative mathematician like Dodgson, since non-commutative algebras contradicted the basic laws of arithmetic and opened up a strange new world of mathematics, even more abstract than that of the symbolic algebraists.

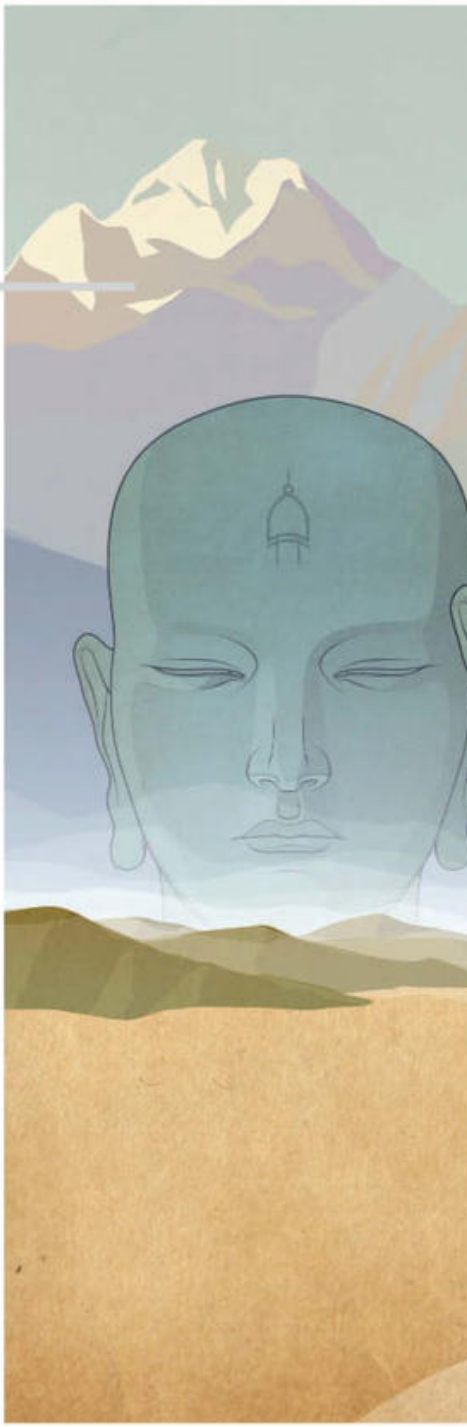
When the scene ends, the Hatter and the Hare are trying to put the Dormouse into the teapot. This could be their route to freedom. If they could only lose him, they could exist independently, as a complex number with two terms. Still mad, according to Dodgson, but free from an endless rotation around the table.

And there Dodgson's satire of his contemporary mathematicians seems to end. What, then, would remain of *Alice's Adventures in Wonderland* without these analogies? Nothing but Dodgson's original nursery tale, *Alice's Adventures Under Ground*, charming but short on characteristic nonsense. Dodgson was most witty when he was poking fun at something, and only then when the subject matter got him truly riled. He wrote two uproariously funny pamphlets, fashioned in the style of mathematical proofs, which ridiculed changes at the University of Oxford. In comparison, other stories he wrote besides the *Alice* books were dull and moralistic.

I would venture that without Dodgson's fierce satire aimed at his colleagues, *Alice's Adventures in Wonderland* would never have become famous, and Lewis Carroll would not be remembered as the unrivalled master of nonsense fiction. ■

God, what a problem

What do you do with the world's hardest logic puzzle? Make it even harder, says *Richard Webb*



RICHARD WILKINSON

HERE'S a puzzle guaranteed to set your grey cells humming.

"Three gods A, B, and C are called, in some order, True, False, and Random. True always speaks truly, False always speaks falsely, but whether Random speaks truly or falsely is a completely random matter. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language in which the words for 'yes' and 'no' are 'da' and 'ja', in some order. You do not know which word means which."

Welcome to the "Hardest Logic Puzzle Ever". If you should happen upon three questions that will unmask the gods, don't stop there. Your next task: make the puzzle even harder.

This is a parlour game played by logicians since the Hardest Logic Puzzle Ever was first so named – and solved – by US logician George Boolos shortly before his death in 1996. Find a solution, and you understand a little more about how to extract truth in a world where imperfect information abounds – and perhaps, by the by, about the nature of logic itself.

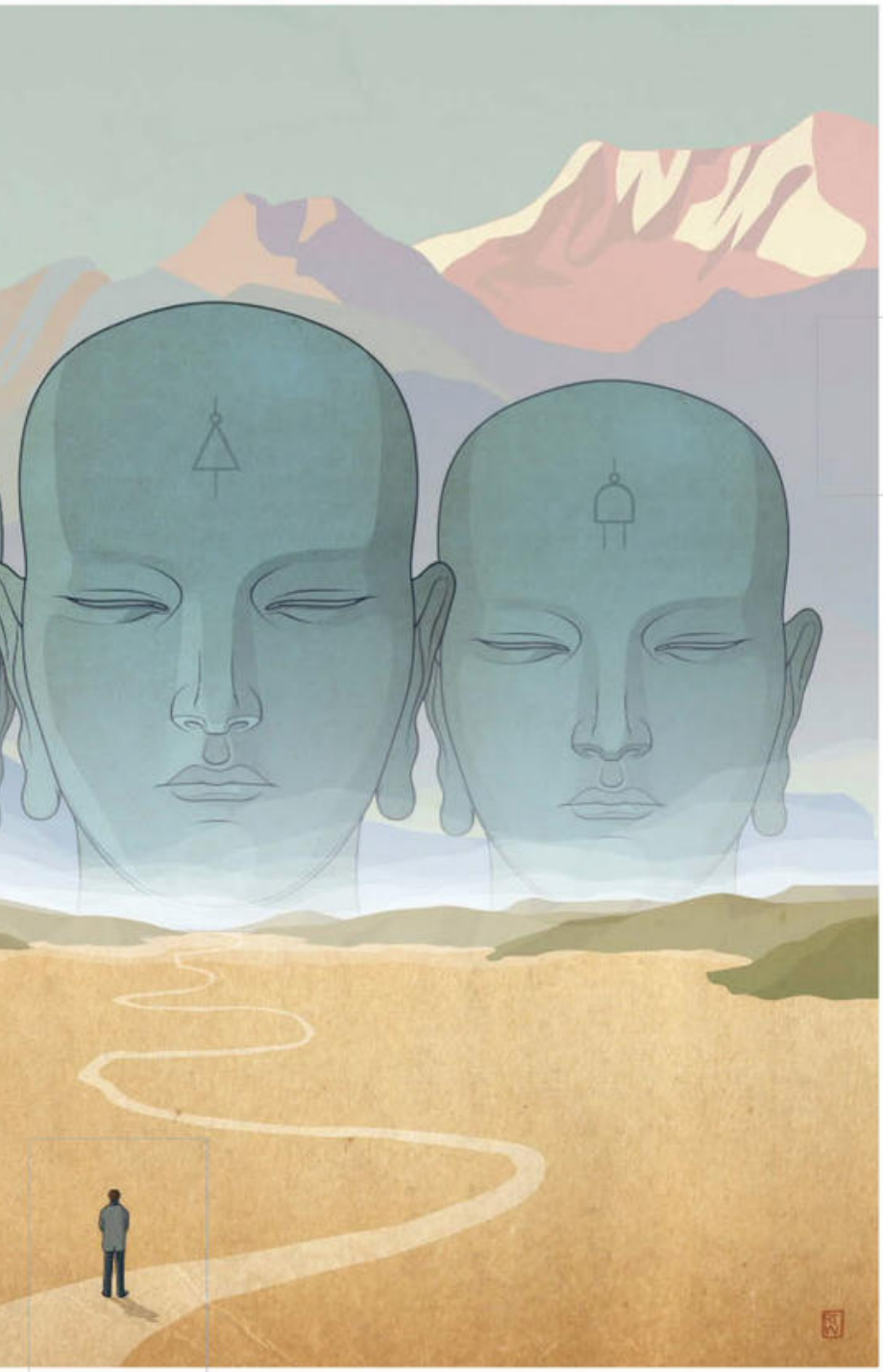
Boolos always had an individual take on the world. He once delivered a public lecture explaining Kurt Gödel's second incompleteness theorem, a seminal result in mathematical logic, entirely in words of one syllable, and was wont to pace the corridors of the Massachusetts Institute of Technology solving problems with invisible chalk on an invisible blackboard he always carried with him. In formulating the Hardest Logic Puzzle Ever, he was building on a series of mind-benders made popular by US mathematician Raymond Smullyan. In these puzzles, you are marooned on an island among knights, who always speak the truth, and knaves, who do nothing but lie. You generally have one question to extract some vital piece of information from them (see diagram, page 124).

Boolos's genius was to compact into one puzzle so many stumbling blocks that only a fiendishly complex series of questions can lead to the solution. "What makes it hard is the combination: liars and truth-tellers, language ignorance and finally a random element," says Brian Rabern, a philosopher at the University of Illinois at Urbana-Champaign. To reproduce Boolos's full answer, which he set out in *The Harvard Review of Philosophy*, would be to spoil the fun. For those needing a head start: his first question, addressed to god A, is, "Does 'da' mean 'yes' if and only if you are True if and only if god B is Random". Now get out those invisible chalkboards.

Boolos's intention in formulating the puzzle was not entirely frivolous. His solution stood or fell on extensive use of one of three classical axioms of logic attributed to Aristotle. Known as the law of excluded middle, it states that a logical proposition must be either true or false; there is no third way.

But is the law of the excluded middle itself true (and if not, is it false)? Consider, for example, the statement, "The present King of America has a beard". Is it necessarily false by virtue of there being no King of America, or does it lie in some grey zone between truth and falsehood? With his solution to the Hardest Puzzle, Boolos aimed to show how difficult it becomes to solve logical problems if one allowed such a middle way.

His solution was not to everyone's



taste. As Tim Roberts of Central Queensland University in Bundaberg, Australia, observed tartly in 2001, the obfuscating "if and only if" statements with which Boolos laced his solution were "the sort of thing that makes most laymen despair of logicians". Producing a solution that did away with them, Roberts concluded that the Hardest Puzzle was not so hard after all, and went on to suggest two more troublesome alternatives: make two gods Random, and the third either True or False;

"George Boolos was wont to pace the corridors solving problems with invisible chalk on an invisible blackboard"

or one god Random and the other two indeterminately either True or False.

The floodgates really opened in 2008, though, when Brian Rabern and his brother Landon discovered a more fundamental flaw in Boolos's original puzzle. It lay in his clarification of how Random generates his answers: like flipping a coin, Boolos specified, where heads makes him speak the truth and tails forces him to lie. In that case, said the Rabern brothers, just ask the question, "Are you going to answer this question with a lie?". True and False can only answer this question with the word meaning "no". If Random's coin shows heads, meanwhile, he must speak truly and also say the word for "no". Equally, if it shows tails, he must lie – again answering in the negative. So whoever you are speaking to, you now know how to say "no" in the gods' language.

Head-exploders

This allows the problem to be solved in three surprisingly easy steps. That isn't all: similarly self-referential questions can also throw True and False into utter confusion. For example, ask them, "Are you going to answer 'ja' to this question?". If "ja" means "no" True cannot say the truth, and if it means "yes" False cannot say a lie, so one or other of them will be left lost for words. "We called them head-exploding questions," says Brian Rabern.

Such undefined statements are the bane of unwary computer programmers, producing a program that is paralysed by indecision. But the Rabern brothers showed how using these statements judiciously unmasked True and False quicker and helped to solve the puzzle in just two questions. Even when Random's behaviour was tweaked to make him answer truly randomly, the puzzle was easily solvable in three steps.

And so it went on. While the validity of head-exploding questions remains questioned (see "Explosive logic", page 124), in 2010 philosopher Gabriel Uzquiano of the University of Oxford embedded them in more complex logical structures to show that you could also solve the truly random version of the ›

puzzle with just two questions – and then came up with a harder variant in which Random could randomly decide to say nothing at all. Later that year, Gregory Wheeler of Carnegie Mellon University in Pittsburgh, Pennsylvania, and his colleague Pedro Barahona responded with a solution to Uzquiano's problem in three questions. A still harder puzzle could be formulated, they suggested, by replacing Random with Devious, who lies when he can but if he gets confused acts like Random.

At the moment, they have their peers stumped with this version. "We have seen some papers come through, but nothing has quite got there," says Wheeler.

So what more is there to this, beyond logical one-upmanship? Quite a bit. "It is not just about logic, it is about information extraction, learning about nature when she is unwilling to give up her secrets," says Wheeler. Brian Rabern agrees. "The god Random makes it a toy model of reasoning with imperfect information, which we must do all the time in normal life," he says. By clarifying how we do that most efficiently, the puzzle hones our logical arsenal – an understanding that could help us to program artificial intelligences to reason about the world.

That thought inspired Nikolay Novozhilov, a hobby puzzler based in Singapore, to bind our hands even tighter. In 2012, he modified the puzzle's set-up so you are given no clues about the gods' language. This means we cannot ask questions such as, "If I asked you X, would you answer 'da'?". "The idea was to find out, if you eliminate all understanding, is it still solvable?" he says.

The answer is yes, provided whoever is being questioned has developed distinct ways to express basic logical concepts such as true and false. That result feeds into a long-running debate between linguists as to the minimum requirement to build a lexicon of a completely unknown language. Novozhilov playfully suggests one situation in which it might come in useful: that first encounter with the little green men. "I am sure that aliens will have the same understanding of logic whatever world they live in," he says. "Even if you don't have any information about how someone communicates, this shows there are features you can predict just from a logical understanding of what language is about."

Perhaps that is an unnecessarily unnerving suggestion with which to justify some fireside puzzle-solving. Wheeler suggests we need not look so far afield for a motivation. "There is something aesthetically lovely about a well-made puzzle." ■

"One situation in which this might come in useful would be our first encounter with the little green men"

EXPLOSIVE LOGIC

Solving George Boolos's "Hardest Logic Puzzle Ever" with "head-exploding" questions that have no true or false answer (see main story) puts the spotlight on the "law of non-contradiction". This axiom of classical logic states that no proposition may be both true and false. Graham Priest of the City University of New York is one logician who thinks it is at best a half-truth. He has spent the past three decades developing "paraconsistent" logical systems that admit the existence of dialetheia, or true contradictions. The initial motivation was to get around the liar paradox – the 2500-year-old unsolved conundrum of what truth is in the statement "this statement is false". "If you're using a paraconsistent logic, you can tolerate that sort of contradiction without it causing havoc elsewhere," says Priest. Some statements simply are both true and false.

Allowing some elasticity in our logic might help us to model the world better under certain circumstances: in quantum physics, for example, where things are not necessarily always one thing or the other, but sometimes a bit of both. It is an approach that dismays purists. "Many theorists wouldn't like the idea of logic being held hostage to the empirical realm," says Brian Rabern, a philosopher at the University of Edinburgh, UK.

That's a debate unlikely to produce a true or false answer soon. With the Hardest Puzzle, allowing a god to be a dialetheist unfazed by head-exploding questions shakes things up once again. "How do you solve the problem then?" asks Priest. "I've absolutely no idea, but it does ratchet things up a notch, which is nice."

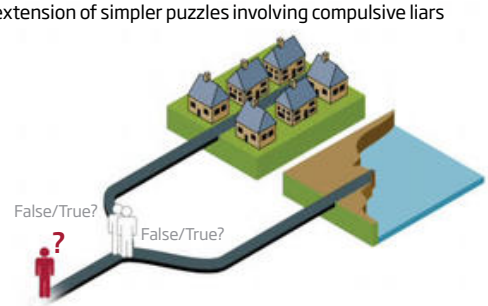
The world's hardest puzzle – easy version

George Boolos's "Hardest Logic Puzzle Ever" is an extension of simpler puzzles involving compulsive liars and inveterate truth-tellers

THE PROBLEM

A fork in a road is guarded by a liar (**False**) and a truth-teller (**True**) – you don't know which is which

What single question, demanding a yes or no answer, can you ask to find out which road leads to the village, and which over the cliff?



THE SOLUTION

Ask the question:

"If I ask the other person if the left path leads to the village, what would he say?"

	You ask True	**You ask False**
If the left path goes to the village	False would say no, so True says NO	True would say yes, so False says NO
If the right path goes to the village	False would say yes, so True says YES	True would say no, so False says YES

THE ANSWER

Regardless of who True and False are, the answer is: NO = LEFT PATH and YES = RIGHT PATH

The real answer to life, the universe and everything

Christopher Kemp discovers the number encoded in our genes

VIJAY MAHRRA/EYEEM/GETTY

MAXIM MAKUKOV has an idea. It's unorthodox; you might call it "out there". Makukov understands that. He knew he'd have his critics the moment he began to develop it. But it's there in the numbers, he says. And numbers don't lie.

A cosmologist and astrobiologist at the Fesenkov Astrophysical Institute in Almaty, Kazakhstan, Makukov says the numbers reveal that all terrestrial life came from outer space. Not only that, it was planted on Earth by intelligent aliens. Billions of years ago, the planet was barren and lifeless. But then, at some distant and unknowable moment, it was seeded with what Makukov calls an "intelligent-like signal" – a signal that is too orderly and intricate to have occurred randomly.

This signal, he says, is in our genetic code. Highly preserved across cosmological timescales, it has been waiting there, like an encrypted message, for anyone qualified to read it. All of the teeming varieties of life on Earth – from kangaroos and daffodils to albatrosses and us – carry it within them. And now Makukov, along with his mentor, mathematician Vladimir *sh*Cherbak of the al-Farabi Kazakh National University in Almaty, claims to have cracked it. If they are right, the answer to life, the universe and everything is... 37.

The idea that terrestrial life has extraterrestrial origins has a long and sometimes distinguished history. The standard version goes something like this: a primitive alien life form, perhaps a bacterium, somehow hitches a ride through space aboard an object like a meteoroid, collides with our young planet and seeds it with life. Against innumerable odds, its descendants flourish and spread across Earth.

In 1871, Lord Kelvin hypothesised "that there are countless seed-bearing meteoric stones moving about through space". In his 1908 book *Worlds in the Making*, Nobel laureate Svante Arrhenius named the process "panspermia". As recently as 2009, Stephen Hawking speculated that "life could spread from planet to planet, or from stellar system to stellar system, carried on meteors".

Prestigious backers notwithstanding, panspermia has not found widespread acceptance, although many biologists accept a weaker version of it. "Most biologists will agree there is a contribution to the origin of life on Earth from cosmic sources," says P. Z. Myers of the University of Minnesota, Morris. "We have lots of organic compounds floating around in space."

Makukov and *sh*Cherbak have taken it further. They're reviving something called "directed panspermia", the hypothesis that life was seeded intentionally by an extraterrestrial intelligence.

The idea goes back to 1973, when Francis Crick published a paper in the planetary sciences journal *Icarus*, at that time edited by Carl Sagan. In it, Crick asked the question: "Could life have started on Earth as a result of infection by microorganisms sent here deliberately by a technological society on another planet, by means of a special long-range uncrewed spaceship?"

Extraordinary claims like this ›

require extraordinary evidence. For more than a century, people have been trying to find at least some of that evidence – proof of the existence of sentient aliens.

The bulk of this effort – known as SETI, or the search for extraterrestrial intelligence – has involved trying to detect radio signals. But despite almost a century of vigilance, says SETI senior astronomer Seth Shostak, they have heard nothing.

With one possible exception. In 1977, SETI researchers in Ohio picked up a 72-second burst of radio waves that was

> "The genetic code is a perfect place to plant a secret message"

so close to what they had been looking for that one of the researchers wrote "Wow!" on the readout. Nothing like the Wow! signal has ever been seen since.

The radio silence has inspired some to widen the search. Many have asked: what if the message is here on Earth already? What if *we* are the message?

In his 2010 book *The Eerie Silence*, Paul Davies, a physicist at Arizona State University, wrote about genomic SETI – the idea that our genome might house a secret message. He was following the physicist George Marx, who in 1979 wrote: "It is possible that a few billion years ago an advanced civilization prepared some sort of message using genetic engineering and sent it to Earth. This extraterrestrial DNA molecule became the starting point of biological evolution."

Makukov and *sh*Cherbak's ideas are in this tradition. But instead of rummaging through DNA, they look to the genetic code, a complex set of rules by which DNA is translated into proteins (see "Code within a code", above right). The genetic code shouldn't be confused with the genome, which is a specific set of genetic instructions for making a fruit fly, say, or a giant redwood. Instead, it specifies how to convert those instructions into proteins.

Unlike genomic DNA, the code is stable. Genomes mutate over time, but the code is passed down the generations without alteration and appears to have remained almost completely unchanged for billions of years.

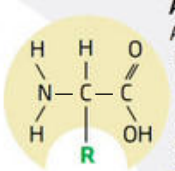
For that reason, says Makukov, it is the

Code within a code

The genetic code consists of the four DNA bases, A, C, G and T, organised into three-letter "codons". There are 64 codons in total; each specifies one of the 20 amino acids that are used to build proteins, or is a stop signal. Some amino acids are specified by one codon, others by up to six.

The code is subdivided into 16 families; each of the four members within a family start with the same two letters. Half of the families are whole families – all four codons code for the same amino acid. The other half are split families – the four codons do not all code for the same amino acid

Whole families
Split families
MM = molecular mass



AMINO ACID
All 20 amino acids specified by the code have a common core but different side chains. Some of the curious patterns in the code are revealed by the molecular masses of the amino acids and their side chains

MM of the core = 74 (= 2×37)
The molecular masses of the side chains range from 1 (glycine, which has a single hydrogen atom) to 130 (tryptophan)

CODON	AMINO ACID	CODON	AMINO ACID	CODON	AMINO ACID	CODON	AMINO ACID
TTT	Phenylalanine (Phe) (MM 165)	TCT	Serine (Ser) (MM 105)	TAT	Tyrosine (Tyr) (MM 181)	TGT	Cysteine (Cys) (MM 121)
TTC		TCC		TAC		TGC	
TTA	Leucine (Leu) (MM 131)	TCA		TAA	Stop	TGA	Stop
TTG		TCG		TAG	Stop	TGG	Tryptophan (Trp) (MM 204)
CTT	Leucine (Leu) (MM 131)	CCT	Proline (Pro) (MM 115)	CAT	Histidine (His) (MM 155)	CGT	Arginine (Arg) (MM 174)
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	Glutamine (Gln) (MM 146)	CGA	
CTG		CCG		CAG		CGG	
ATT	Isoleucine (Ile) (MM 131)	ACT	Threonine (Thr) (MM 119)	AAT	Asparagine (Asn) (MM 132)	AGT	Serine (Ser) (MM 105)
ATC		ACC		AAC		AGC	
ATA		ACA		AAA	Lysine (Lys) (MM 146)	AGA	Arginine (Arg) (MM 174)
ATG	Methionine (Met) (MM 149)	ACG		AAG		AGG	
GTT	Valine (Val) (MM 117)	GCT	Alanine (Ala) (MM 89)	GAT	Aspartic acid (Asp) (MM 133)	GGT	Glycine (Gly) (MM 75)
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	Glutamic acid (Glu) (MM 147)	GGA	
GTG		GCG		GAG		GGG	

Symmetries of 37

37×1 037	37×2 074	37×3 **111**	37×4 148	37×5 185	37×6 **222**	37×7 259	37×8 296	37×9 **333**
37×10 370	37×11 407	37×12 **444**	37×13 481	37×14 518	37×15 **555**	37×16 592	37×17 629	37×18 **666**
37×19 703	37×20 740	37×21 **777**	37×22 814	37×23 851	37×24 **888**	37×25 925	37×26 962	37×27 **999**

481 ÷ 37 = 13 = 4+8+1
629 ÷ 37 = 17 = 6+2+9
777 ÷ 37 = 21 = 7+7+7 etc

Rumer's transformation

Whole families can be converted into split families – and vice versa – by switching Ts for Gs and As for Cs

The probability of this happening by chance is extremely small

TT	⟷	GG
TG	⟷	GT
TA	⟷	GC
TC	⟷	GA
GG	⟷	TT
GT	⟷	TG
GA	⟷	TC
GC	⟷	TA
AA	⟷	CC
AT	⟷	CG
AG	⟷	CT
AC	⟷	CA
CC	⟷	AA
CT	⟷	AG
CG	⟷	AT
CA	⟷	AC

Transformation #2

Another pattern emerges using Rumer's transformation on the subset of codons with either three identical or three different bases

Phe 165	TTT	⟷	GGG	Gly 75
Lys 146	AAA	⟷	CCC	Pro 115
Ile 131	ATC	⟷	CGA	Arg 174
Ser 105	TCA	⟷	GAC	Asp 133
His 155	CAT	⟷	ACG	Thr 119
Ala 89	GCT	⟷	TAG	Stop 0
Leu 131	CTG	⟷	AGT	Ser 105
Cys 121	TGC	⟷	GTA	Val 117
Stop 0	TGA	⟷	GTC	Val 117
Asp 133	GAT	⟷	TCG	Ser 105
Met 149	ATG	⟷	CGT	Arg 174
Gln 146	CAG	⟷	ACT	Thr 119
Ser 105	AGC	⟷	CTA	Leu 131
Ala 89	GCA	⟷	TAC	Tyr 181
1665 (=999 + 666)	*Total molecular mass*			1665 (=999 + 666)
703 (=19×37)	*Total mm of side chains*			703 (=19×37)

The pattern is revealed in the combined molecular masses of the amino acids on either side of the transformation

perfect place to plant a message. Billions of years ago, he says, that is precisely what happened.

To test the idea, Makukov and *sh*Cherbak devised a mathematical approach to analyse the code, searching for patterns unlikely to occur at random.

Their arguments are often dense and impenetrable, filled with complex mathematical formulae. But at heart, Makukov says, "it's very simple". The genetic code is like some type of combinatorial puzzle, he says. In other words, once you begin to analyse it, hidden regularities emerge.

"It was clear right away that the code has a non-random structure," says Makukov. "The patterns that we describe are not simply non-random. They have some features that, at least from our point of view, were very hard to ascribe to natural processes."

Exhibit A is Rumer's transformation. In 1966, Soviet mathematician Yuri Rumer pointed out that the genetic code can be divided neatly in half (see "Rumer's transformation", left). One half is the "whole family" codons, in which all four codons with the same two initial letters code for the same amino acid. The AC family, for instance, is "whole" because codons beginning AC code for threonine. On the other are "split family" codons, which don't have this property.

Rumer first noted that there is no good reason why exactly half of the codons should be whole. More profoundly, he also realised that applying a simple rule – swapping T for G, and A for C – converts one half of the code into the other.

That might sound inevitable, but it is not. In 1996, mathematician Olga Zhaksybayeva of the al-Farabi Kazakh National University calculated that the probability of it occurring by chance is 3.09×10^{-32}.

And Rumer's transformation is just one of many patterns and symmetries within the code. Another example: you can create a subset of codons including those with three identical bases (AAA, say) and those with three unique bases (GTC, say). Using a Rumer-type transformation, these 28 codons can be divided into two groups each with a combined total atomic mass of 1665, and a combined "side chain" atomic mass of 703 (see "Transformation #2", left). Both are multiples of the prime number 37, which has interesting mathematical properties of its own (see "Symmetries of 37", opposite).

In fact, 37 recurs frequently in the code. For example, the mass of the molecular "core" shared by all 20 amino acids is 74, which is 37 doubled. Forget 42...

All in all, the Kazakhs have identified nine patterns in the code, which they spell out in detail in a paper published in 2013 under the provocative title "The 'Wow! signal' of the terrestrial genetic code".

If you think that all sounds a bit like *The Da Vinci Code for DNA*, you're not alone. "It's flat out numerology," says Myers, who also notes the similarity to the pseudoscience of intelligent design – a comparison Makukov and *sh*Cherbak reject. "The hypothesis has nothing to do with intelligent design," they say.

Others are less critical. "It's not, in and of itself, absurd," says David Grinspoon, senior scientist at the Planetary Science Institute and author of *Lonely Planets: The natural philosophy of alien life*. "We're already learning to custom design organisms and we're already learning to send things out into space. If anybody else is out there, the chances are they're not as new at it as we are."

> "What is the probability of finding something like this by chance?"

Davies is also quite forgiving. "If you crunch numbers long enough, you'll find patterns in almost anything," he says. "It was very clear to me at the outset that what this boils down to is an assessment: what is the probability that you might find something like this by chance?"

To that, Makukov and *sh*Cherbak have an answer: about 10^{-13}, or 1 in 10 trillion. In 2014, they published a second paper on the work.

As to what – or who – planted the message, Makukov stresses that he doesn't know. "This is speculation," he says. "Maybe they're gone long ago. Maybe they're still alive. I think these are questions for the future."

But on the basic idea, he is adamant. "For the patterns in the code," says Makukov, "the explanation we give, we think is the most plausible." ■